Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex

Allopolyploidy combines two progenitor genomes in the same nucleus. It is a common speciation process, especially in plants. Deciphering the origins of polyploid species is a complex problem due to, among other things, extinct progenitors, multiple origins, gene flow between different polyploid populations, and loss of parental contributions through gene or chromosome loss. Among the perennial species of *Glycine*, the plant genus that includes the cultivated soybean (*G. max*), are eight allopolyploid species, three of which are studied here. Previous crossing studies and molecular systematic results from two nuclear gene sequences led to hypotheses of origin for these species from among extant diploid species. We use several phylogenetic and population genomics approaches to clarify the origins of the genomes of three of these allopolyploid species using single nucleotide polymorphism data and a guided transcriptome assembly. The results support the hypothesis that all three polyploid species are fixed hybrids combining the genomes of the two putative parents hypothesized on the basis of previous work. Based on mapping to the soybean reference genome, there appear to be no large regions for which one homoeologous contribution is missing. Phylogenetic analyses of 27 selected transcripts using a coalescent approach also are consistent with multiple origins for these allopolyploid species, and suggest that origins occurred within the last several hundred thousand years.

Aureliano Bombarely[1], Jeremy E. Coate[2], and Jeff J. Doyle[1]

1 - Department of Plant Biology, Cornell University, Ithaca, New York 14853 USA; 2 - Department of Biology, Reed College, Portland, Oregon 97202-8199 USA;

INTRODUCTION

Polyploidy (whole genome duplication, WGD) is a key process in plant evolution. All seed plants are fundamentally polyploid, with a second WGD event shared by all flowering plants (Jiao et al., 2011), and additional events found in many lineages (see http://genomevolution.org/wiki/index.php/Plant_paleopolyploidy)(Soltis et al., 2009). It has been estimated that 15% of all flowering plant speciation events involve polyploidy (Wood et al., 2009). Systematists generally recognize autopolyploidy and allopolyploidy as distinct types of polyloidy events, based on the level of divergence of the diploid genomes that formed the polyploid. The terms are best thought of as describing elements of a continuum that ranges from the doubling of a single genome (autopolyploidy), to the incorporation of differentiated genomes in a single nucleus by hybridization of different species (allopolyploidy). From a genetic perspective, allopolyploids are characterized by diploid-like meiotic behavior and limited interaction between the two homoeologous genomes. The duplicated chromosomes of an autopolyploid (and, to a lesser extent, a newly formed allopolyploid; Ramsey and Schemske, 2002) initially can associate randomly, leading to polysomic segregation, but it is generally assumed that this is a transient state; diploidization leads to the eventual presence of homoeologous genomes. It is difficult, if not impossible to determine from the genomes of older polyploids (paleopolyploids, mesopolyploids) how differentiated their progenitor genomes were in large part due to the frequent absence of extant diploid progenitors for comparative purposes.

The initial "fixed hybrid" condition of an allopolyploid erodes over time as homoeologous loci are lost (Lynch and Conery, 2000; Maere et al., 2005); this process of "fractionation" is thought to occur preferentially from one subgenome, but the precise mechanisms remain unknown (Schnable and Freeling, 2011; Freeling et al., 2012). In addition to the loss of genes, the process of concerted evolution can result in the replacement of a gene from one genome by its homoeologue, notably through gene conversion (e.g., Wang et al., 2007). The earliest stages of polyploid evolution may contribute disproportionately to gene loss and genomic rearrangement through genomic shock (McClintock, 1984). For example, some individuals of the ca. 100 year-old allopolyploid, *Tragopogon miscellus*, have lost entire chromosomes of one parent (Chester et al., 2012). Diversity in polyploids can be due to mutational divergence from parental diploids, but also due to multiple origins produced by different polyploidization events between different genotypes of the same diploid species (Symonds et al., 2010). Questions concerning how polyploids originate (e.g., single vs. multiple origins), how they partition their variation (e.g., as a single lineage united by gene flow vs. as separate lineages formed from different genotypes of the same progenitor species), and how much of the initial parental contributions they retain are among the major questions in polyploid evolutionary research (Soltis et al., 2010).

High-throughput sequencing produces massive amounts of genome-wide data, and thus has great potential for systematic and evolutionary studies in general (Gilad et al., 2009). The ready availability of genomic and transcriptomic data has opened new opportunities for studying the evolution of polyploids (Bombarely et al., 2012; Grover et al., 2012; Ilut et al., 2012; Dufresne et al., 2014) at the scale of whole genomes. However, it is not trivial to extract relevant information from short read sequencing data, particularly for allopolyploids, where the interest is often in deconvoluting the complex genome into its two homoeologous subgenomes (Grover et al., 2012; Ilut et al., 2012). Moreover, the field of systematics has what has been called a new paradigm for studying species relationships, involving genealogical approaches (Edwards, 2009). Genealogical methods have lately begun to be applied to both autopolyploids (Arnold et al., 2012; Hollister et al., 2012) and allopolyploids (e.g. Slotte et al., 2011; Jones et al., 2013; Slotte et al., 2013). The confluence of these two developments promises to accelerate the study of polyploid evolution.

The genus *Glycine* includes the cultivated soybean (*G. max*) and its wild progenitor (*G. soja*), both annual species native to northeastern Asia, as well as approximately 30 perennial species native to Australia classified as subgenus *Glycine* (Ratnaparkhe et al., 2011). Like many plant species, *Glycine* has a complex history of polyploidy: in addition to events shared with all angiosperms (Jiao et al., 2011) and eudicots (Jiao et al., 2012), the soybean

genome retains evidence from a WGD around 50 million yeas ago (MYA) shared with a large subset of legumes (Blanc and Wolfe, 2004; Schlueter et al., 2004; Cannon et al., 2010), and particularly from a more recent polyploidy event that increased the chromosome number of the ancestor of all extant *Glycine* species from $2n = 20$ to $2n = 40$ (Shoemaker et al., 2006; Doyle and Egan, 2010; Schmutz et al., 2010; Doyle, 2012). This *Glycine*-specific WGD occurred between the estimated time of homoeologous gene divergence in the soybean genome (10-13 MYA; e.g., Egan and Doyle, 2010; Schmutz et al., 2010), and around 5 MYA, when the annual and perennial species diverged from an already-polyploid common ancestor (Doyle and Egan, 2010).

In addition to these older events, eight perennial *Glycine* species are allopolyploids with $2n = 78$ or 80, hypothesized to have arisen by hybridization involving various combinations of eight extant diploid species, several of them multiple times and involving both progenitors as chloroplast genome donors (Doyle et al., 2004). Various lines of evidence culminated in these hypotheses of reticulate relationships, which are shown in Fig. 1 for the six species that are part of the *G. tomentella* sub-complex (Doyle et al., 2004). Chromosome number polymorphism ($2n = 38, 40, 78, 80$) was observed in what was initially considered a single taxon, *Glycine tomentella* (Newell and Hymowitz, 1978). Patterns of sterility and partial chromosome pairing in artificial crosses among *G. tomentella* plants were consistent with the presence of shared homoeologous diploid genomes among polyploids (Grant et al., 1984; Doyle et al., 1986; Singh et al., 1988). Isozyme studies of diploid and allopolyploid *G. tomentella* led to the characterization of numerous "races" designated either "D" for diploid, or "T" for tetraploid (Doyle et al., 1985; Singh et al., 1998). Morphological complexity, presumably due to the reticulate nature of the complex, has slowed nomenclatural recognition of what are clearly species in the biological sense. More recently, molecular phylogenetic studies assumed a dominant role in refining hypotheses of relationships (Hsing et al., 2001; Singh et al., 1998; Brown et al., 2002; Doyle et al., 2002; Rauscher et al., 2004), and corroborated earlier hypotheses concerning the origins of polyploids from among the diploid ($2n = 38, 40$) "genome groups" that were also initially defined by artificial hybridization studies and later by molecular studies (see Ratnaparkhe et al., 2011). However, these DNA-level studies were based on only two molecular markers: the internal transcribed spacers of the 18S-5.8S-26S nuclear ribosomal gene cistron (nrDNA ITS) and the low copy nuclear gene, histone H3D. Relationships of chloroplast genomes are broadly consistent with these results (Hsing et al., 2001), but are complicated by incongruence with nuclear markers, likely due to a combination of incomplete lineage sorting and introgression (Doyle et al., 2004). Thus, a genome-wide perspective on the origin and evolution of the *G. tomentella* complex, including estimates of dates of origin, has been lacking.

A better understanding of the origin and evolution of the *Glycine* allopolyploid complex will complement its exploitation in studying the impact of allopolyploidy on a range of morphological and physiological characters (Coate and Doyle, 2010; Coate et al., 2012; Ilut et al., 2012; Coate et al., 2013; Hegarty et al., 2013). Here we apply phylogenetic and coalescent methods to a transcriptomic dataset from three of these allopolyploid species and their diploid progenitors that was originally generated to study the effects of polyploidy on their ability to cope with stress from excess light (Coate et al., 2013).

MATERIAL AND METHODS

Taxon Sampling and Transcriptome Sequencing

Three *Glycine* (Fig. 1) allopolyploid "triads" (from the *Glycine* perennial polyploid complex) defined as an allopolyploid species and its two putative diploid progenitors, were sampled: 1) the allopolyploid, *G. tomentella* T1 ($2n = 78$) and the diploid species, *G. tomentella* D1 (E-genome of Hymowitz et al., 2010; $2n = 38$) and *G. tomentella* D3 (D-genome; $2n = 40$); 2) *G. dolichocarpa* (= *G. tomentella* T2; $2n = 80$) and its putative progenitors *G.*

tomentella D3 and *G. syndetika* (= *G. tomentella* D4; A-genome; 2*n* = 40); and 3) *G. tomentella* T5 (2*n* = 78) and its hypothesized progenitors, *G. tomentella* D1 and *G. clandestina* (A-genome; 2*n* = 40). Each species was represented by 2-5 accessions sampled from the CSIRO Division of Plant Industry Perennial *Glycine* Germplasm Collection (Table 1). Additionally, a synthetic allotetraploid (A58) was used, which mimics the natural T5 alopolyploid, having been produced by doubling an artificial hybrid of *G. tomentella* D1 (accession G1316) and *G. canescens* (accession G1233; A-genome; 2*n* = 40); *G. canescens* is an A-genome species closely related to *G. clandestina.* A summary of the datasets used can be found in Suplementary Table 4.

Plants were grown in a common growth chamber with a 12 h/12 h light/dark cycle, 22 ºC/18 ºC day/night temperature regime, and a light intensity of either 125 mmol m^{-2} s^{-1} (LL) or 800 mmol m^{-2} s^{-1} (EL). Different light intensities were used for the purposes of a separate study examining light stress responses (Coate et al., 2013). Single leaflets were pooled from six individuals per accession, and RNA-Seq libraries were constructed from the pooled tissue. All samples were taken from approximately 1-week-old, fully expanded leaves, and were collected 0.5–2.0 h into the light period. For each light treatment, all tissue was collected in a single morning and immediately frozen in liquid nitrogen. Total RNA was isolated from pooled leaf tissue using the Plant RNeasy Kit with on-column DNase treatment (Qiagen, Valencia, CA, USA). Single-end RNA-Seq libraries were constructed following the Illumina mRNA-seq Sample Preparation Kit protocol (Illumina, San Diego, CA, USA), with the following modifications: (1) two rounds of polyA selection were performed using the Dynabeads mRNA DIRECT Kit (Life Technologies, Carlsbad, CA, USA); (2) RNA was fragmented for 2 min at 70 ºC using the RNA fragmentation reagents kit (Life Technologies; and (3) Illumina PE adapters were replaced with custom-made adapters containing 3nt barcodes in order to facilitate multiplexing of samples (see Coate et al., 2013 for adapters and Supplementary Table S1 for the barcode sequences). Sequencing was performed on either the GAIIx or HiSeq 2000 platform (Illumina), generating 88 nt or 100 nt reads, respectively. Equimolar amounts of three (GAIIx) or four (HiSeq 2000) barcoded libraries were combined and sequenced per channel.

Read Processing and Single Nucleotide Polymorphism (SNP) Calling

Reads were processed with Fastq-mcf (Aronesty, 2013) to trim low quality extremes (min. quality 30) and remove short reads (min. read length 50 bp). They were aligned to the soybean genome (version 1.0, downloaded from www.phytozome.net/soybean) using Bowtie2 (Langmead and Salzberg, 2012) with the default parameters. Mapping files from the same accession were merged. Reads without preferential mapping (same score for two or more mapping hits) and with a mapping score below 20 were removed. SNP calling was performed using Samtools (Li et al., 2009). SNPs supported with read coverage below 5 were removed. VCF files were combined and formatted to Structure and Hapmap formats using the Perl script MultiVcfTool (https://github.com/aubombarely/GenoToolBox/blob/master/SeqTools/MultiVcfTool).

Homoeologue read identification and transcript-guided assembly

For homoeologous SNP identification, a consensus diploid transcriptome was rebuilt for each of the species groups (A, with *G. clandestina* and *G. canescens* accessions; D1, with *G. tomentella* D1 accessions; D3, with *G. tomentella* D3 accessions; and D4, with *G. syndetika* accessions) using Samtools (Li et al., 2009) and Gffread from the Cufflinks software package (Trapnell et al., 2010). A progenitor reference set was created for each of the polyploid species joining the diploid transcriptome sets (T1=D1+D3, T2=D3+D4 and T5=A+D1). Reads from the polyploid species were mapped with these references using Bowtie2. Sam mapping files were processed to identify reads according the preferential mapping with each of the progenitors using the Perl script, SeparateHomeolog2Sam (https://github.com/aubombarely/GenoToolBox/blob/master/SeqTools/SeparateHomeolog2Sam). Reads with

mapping score AS and XS = 0 (No SNPs) were kept and used to rebuild the polyploid transcriptomes using Samtools (Li et al., 2009) and Gffread (from the Cufflinks package, Trapnell et al., 2012). Once the reads were separated according its preferential mapping, they were mapped back to the soybean genome. SNPs were called as described above.

Population structure analysis

The programs Structure (Pritchard et al., 2000) and fineStructure (Lawson et al., 2012) were used to analyze population structure of the two SNP datasets, with and without polyploid SNPs separated by homoeologue, described above. For Structure, each of the datasets was divided into three subsets of 20,000 SNPs selected with a random function incorporated in the MultiVcfTool. 5 replicates were run for each of the subsets with a burn-in of 10,000 and a number of MCMC repetitions of 10,000, from K=1 to K=15 using the default parameters (λ=1, assuming uniform distribution of allele frequencies, Pritchard et al., 2000). Admixture was selected.. The optimal number of clusters was identified based on the rate of change in the log probability of data between successive K values (Evanno et al., 2005). Results at K=6 were verified with a re-analysis using a burn-in of 100,000 generations. Results were visualized using R (barplot function).

The two SNP datasets were divided into 20 different subsets each mapping to one soybean reference chromosome for FineStructure analysis. Analyses were performed following the instructions from the fineStructure web for the unlinked model (http://www.maths.bris.ac.uk/~madjl/finestructure/data_example.html). Results were presented as a heatmap of distances between each of the accessions. A principal component analysis (PCA) was performed over the same distance matrix using fineStructure software. The PCA figure was created using R.

Reconstruction of phylogenies using concatenated SNPs

SNPs from the dataset in which SNPs from allopolyploids were partitioned into their two homoeologues ("homoeologue data set") and were concatenated to create a supermatrix with 36 operational taxonomic units (OTUs). The two homoeologous gene copies from each allopolyploid were treated as individual OTUs; for example the D1 and D3 homoeologues of T1 individuals were treated as D1T1 and D3T1, respectively. *G. max*, accession William82 was used as outgroup. The alignment files were produced changing the SNPs Hapmap format to fasta using a Perl script. The resulting matrix was used in two analyses. First, maximum likelihood (ML) was used, implemented in PhyML (Guindon and Gascuel, 2003) with GTR as the substitution model; 100 bootstrap replicates were conducted. Second, in order to visualize reticulations in the dataset, a network method, NeighborNet, was implemented in the SplitsTree package (Huson and Bryant, 2006) with the default parameters. Trees were visualized and drawn using FigTree (Rambaut, 2012).

Gene-based analyses

A subset of transcripts was selected for phylogenetic and network analyses based on the following criteria: No more than 10% of Ns for the guided assembly consensus sequence in any of the accessions after the homoeologue read identification; alignments with at least 1000 bp; and genes with their corresponding *G. max* homologue identified as an existing pair retained from the most recent (ca. 5-10 million years; (Doyle and Egan, 2010)) *Glycine* WGD event, as compiled by Du et al. (Du et al., 2012). Sequence alignments were based on the transcriptome-guided assembly. Sequence for each of the genes was collected with a Perl script (FastaSeqExtract, GenoToolBox script package), concatenated and changed to the required sequence alignment format using a BioPerl script (bp_sreformat.pl). The 95 alignments selected were used in an exploratory phylogenetic analysis using the

Bayesian MCMC method, BEAST (Drummond et al., 2012) (HKY substitution model, 10,000,000 MCMC). Alignments that produced trees in which *G. max* was not sister to perennial *Glycine* species in the consensus tree were removed. Generally the removed alignments showed tree topologies with two large clades with long branches, indicating the possibility of inclusion of paralogous genes from the older whole genome duplication (ca. 50 MY, common to the Leguminosae; reviewed in Doyle 2012) instead the orthologue.

27 genes selected after this filtering were analyzed using three different methods: 1) Phylogenies were reconstructed using ML using PhyML (Guindon and Gascuel, 2003) with 1,000 bootstraps. jModelTest2 was used to choose the best substitution model (Darriba et al., 2012). According to the Bayesian Information Criterion (BIC) HKY was the preferred model (40% of the genes), followed by K80 (26% of the genes; Supplementary Table S2). 2) Networks were constructed using NeighborNet in SplitsTree4 with the default parameters (Huson and Bryant, 2006). 3) Bayesian analysis was performed using BEAST v2.0 (Drummond et al., 2012). The two homoeologous gene copies from each allopolyploid were treated as individual OTUs as in the concatenated analysis, and *G. max*, accession William82 was again used as outgroup. Based on the jModelTest2 results, HKY was used as the substitution model. The MCMC chain was set to 100,000,000 MCMC generations, taking samples every 1000 generations. Divergence ages were estimated by scaling the tree root (divergence between *G.max* and perennials) to 5 Myr (Egan and Doyle, 2010). All trees were drawn using FigTree (Rambaut, 2012).

Species tree reconstruction

Species tree reconstruction under the coalescent was performed using the 27 selected genes in *BEAST (Drummond et al., 2012). The 24 accessions, including two homoeologues for each allopolyploid accession, were grouped in 11 operational taxonomic units (OTUs) for this analysis: *G. canescens, G. clandestina, G. tomentella* D1*, G. tomentella* D3*, G. syndetika (D4), G. tomentella* T1-D1*, G. tomentella* T1-D3*, G. dolichocarpa* T2-D3*, G. dolichocarpa* T2-D4*, G. tomentella* T5-A *and G. tomentella* T5-D1. *G. max* was used as outgroup. Based on jModelTest2 results, HKY was used as substitution model. The MCMC chain was set to 100,000,000 MCMC generations, taking samples every 1000 generations. Divergence dates were estimated as described above. All the trees were drawn using FigTree (Rambaut, 2012).

RESULTS

Phylogenomics dataset generation

Between 7-60 million reads from leaf transcriptomes of 24 accessions representing 8 *Glycine* perennial species were mapped to the *Glycine max* genome (v1.0) (Schmutz et al., 2010). Reads mapped to 22,500-25,000 genes (~40% of soybean gene models; Table 1); this represents between 4.5 and 11.6% of the genome. 200,000-965,000 single nucleotide polymorphisms (SNPs) were identified relative to *G. max*; 6.3-12.6% of SNP positions were polymorphic in diploid species (*G. clandestina*, *G. canescens*, *G. tomentella* D1 (referred as D1 hereafter), *G. tomentella* D3 (referred as D3) and *G. syndetika* (referred as D4)), and 18.4-28.8% in polyploid species (*G. tomentella* T1 (referred as T1), *G. tomentella* T5 (referred as T5) and *G. dolichocarpa* T2 (referred as T2); Table 2). The interpretation of these positions as standard heterozygosity is complicated by the recent (5-10 MYA: Doyle and Egan, 2010) WGD in the ancestral *Glycine* genome. In a gene for which soybean has lost one of the homoeologous copies from this event, but the perennial species for which it is serving as reference has retained both copies, polymorphic SNPs may be due to reads from two different homoeologous loci in the perennial, rather than two alleles at a single locus. Low levels of conventional heterozygosity are expected in *Glycine* species, because of their strongly selfing reproductive biology, with much reproduction occurring through cleistogamous (closed, selfing)

flowers.

The much higher percentage of polymorphic positions in polyploid individuals (T1, T2, T5) likely is also due to the mapping of reads from two homoeologous copies to a single target, in this case due to recent polyploidy: for example, mapping reads from tetraploid ($2n = 80$) T2 to a single locus in the diploid ($2n = 40$) *G. max* reference genome will result in reads from both its D3 and D4 homoeologous subgenomes mapping to the same target, increasing the chance of observing a polymorphism at a given site. Separating reads from T1, T2, and T5 polyploid individuals was possible where the read has at least one SNP that could be related to one homoeologous genome contributor (e.g., D3 and D4 differed by a SNP and this difference was retained in the D3 and D4 homoeologous genomes of T2; diploid-distinguishing polymorphism (DDP; see Ilut et al., 2012). Between 11.4 and 20.8% of reads were assigned to one of the progenitors (Table 3).

Between 124,984 and 399,884 SNPs were produced for each accession. The filtering of the missing data produced 237,243 and 75,958 polymorphic positions for all the accessions before and after the homoeologous read assignment, respectively. SNPs per chromosome ranged from 7,455 (chromosome 14) to 16,494 (chromosome 8) and from 2,288 (chromosome 14) to 5,300 (chromosome 8) before and after the homoeologous read assignment, respectively.

Transcriptome-guided assemblies produced between ~1,800 and ~6,600 full-length sequences (as mapped to the *G. max* gene models) for each diploid accession. For polyploid subtranscriptomes this number was much lower because only reads that mapped preferentially to one of the diploid consensus species and reads that mapped equally but with no polymorphism (perfect match) were used during the transcriptome-guided assembly. Any read that mapped equally to two or more positions with one or more polymorphisms was discarded because it was impossible to assign it to any of the diploid progenitors, reducing the mapping coverage of the reference gene models. Between ~350 and ~1,350 full length sequences were assembled for the T1, T2, and T5 polyploid homoeologous subtranscriptomes of which between 4 to 19% were duplicated genes from the 5-10 MYA WGD event in the common ancestor of *Glycine* species (Schmutz et al., 2010). For phylogenetic analysis, full length sequences are not needed so a phylogenetic analysis dataset was created with 27 genes (see Material and Methods for the criteria used to generate this dataset; Table 5).

Genome-wide distribution of homoeologous SNPs.

For each allopolyploid accession, the ca. 120,000-400,000 SNPs (Table 3) that could be identified to homoeologous subgenome were mapped to the soybean reference genome (Schmutz et al., 2010). This produced a map that is analogous to chromosome painting (genomic in situ hybridization, GISH) experiments using the reads from which the SNPs were derived, which we term “electronic chromosome painting” (e-chromosome painting). Similar patterns were seen for all accessions, with high densities of SNPs at the ends of each soybean chromosome and far lower densities in pericentromeric regions (Fig. 2). This pattern is expected using reads from transcriptome data, because of the sparse distribution of genes in pericentromeric regions of the soybean genome (Schmutz et al., 2010). Notably, in all allopolyploid accessions, SNPs from both homoeologues were distributed across the entire genome, and no regions were identified in which SNPs from only one homoeologue were mapped (Fig 2; Supplementary Figs. 1-10).

Population structure analyses.

Structure (Pritchard et al., 2000) was first run using all available SNPs, without separating SNPs from polyploids into homoeologous groups. Structure was run from K = 1-15; K = 6 was identified as one of the optimal preferred values of K using the delta K method of Evanno et al. (2005; Supplementary Fig. 11). Five of these six

groups corresponded to diploid taxa: D1, D3, D4, *G. canescens*, and *G. clandestina* (Fig. 3a). The sixth group was represented only as a minor component in D4 accession 2073. Diploid accessions showed little or no evidence of admixture, with the exception of D4 accession 2073 (Fig. 3). In contrast, all polyploid accessions were admixed, each with approximately 50% contributions from two different diploid groups. The genomic makeup of each accession was as expected from previous hypotheses (e.g., Doyle et al., 2002; Fig. 1): T1 accessions showed admixture from D1 and D3, T2 accessions from D3 and D4, and natural T5 accessions from D1 and *G. clandestina*. The synthetic T5 accession (A58) was also admixed, with contributions from D1 and *G. canescens*, as expected (Joly et al., 2004).

A second Structure analysis was conducted with each polyploid accession treated as two separate OTUs, using the homoeologue dataset (Table 2). As with the previous analysis, the analysis was run for K = 1-15. The Evanno method (Evanno et al., 2005) identified K = 6 and 9 as the preferred values (Supplementary Fig. 11). In the case of K = 9 the group representation shows the same structure than the K = 6 (Supplementary Fig. 12). Results for diploids were similar to those obtained in the previous analysis (Fig. 3b). Subgenomes from natural allopolyploids and the synthetic T5 allopolyploid (A58) were shown to belong exclusively to diploid groups, with little or no evidence of admixture, indicating that the SNP filtering into homoeologous contributions was successful.

Complementary to the second Structure analysis, the data were analyzed using ChromoPainter and FineStructure (Lawson et al., 2012). ChromoPainter produces a co-ancestry matrix (as a measure of the ancestry sharing between individuals) based on the haplotype information provided by shared chunks (regions) of biallelic markers between individuals (Lawson et al., 2012). The two SNP datasets were filtered by selecting only the biallelic markers, producing a subset with 220,952 and 71,610 SNPs (before and after homoeologous read assignment, respectively) distributed along all 20 soybean chromosomes. Regions identified by ChromoPainter for each accession ranged from 516 (D4 2321) to 567 (*G. clandestina* 1253) and from 202 (D4 1300 and 2321) to 221 (D4 2073) (before and after homoeologous read assignment respectively). Principal component analysis (PCA) and population relationship analysis using a Bayesian approach were performed over the co-ancestry matrix using FineStructure (Lawson et al., 2012). PCA before homoeologous read assignment (Fig. 4a) shows seven well-differentiated groups, one per species with the exception that *G. canescens* and *G. clandestina* clustered together. Diploid species formed the vertices of a trapezoid. A-genome species (*G. canescens*, *G. clandestina* and D4) formed a more dispersed group than either D1 or D3. Each polyploid species fell between its putative diploid progenitors, consistent with each being an admixture (fixed hybrid). After the homoeologous read assignment (Fig. 4b), each of the polyploid subgenomes clustered with its diploid progenitors, producing three clear clusters: D1, D3, and A-genome (comprising *G. canescens, G. clandestina* and D4, as expected). Heatmaps were used to visualize the population relationships produced by FineStructure, complementing the information shown by the PCA figures. The heatmap before homoeologous read assignment (Fig. 4c), showed four intense regions (red, magenta and blue colors) corresponding to the four species groups of the PCA (Fig. 4a). Each polyploid showed the expected similarity to its progenitors; similarly, as expected the two *G. clandestina* accessions were more similar to one another than either was to *G. canescens*. Also, T5 A58, the artificial polyploid produced from a cross between *G. canescens* 1232 and D1 1316, showed the expected relationships with these accessions. Other T5 polyploids also showed a stronger signal from D1 1316 than from other D1 accessions. T2 accessions did not show any stronger signal with any particular D3 accession than with others, but they did with the D4 accessions 1300 and 2321, relative to 2073. T1 accessions 1288 and 1763 also showed a stronger signal with particular D1 and D3 accessions, whereas T1 accession 1361 showed a weaker signal with the D1 and D3 accessions included here. After the homoeologous read assignment (Fig. 4d), some of these signals were intensified, such as the relationship between T5-D1 subgenomes and particular D1 accessions, but other relationships that were suggested when all SNPs were considered were not observed (for example there is not a stronger signal of D1 1316 with the T5 accessions). These differences may be due to the methodology used for the homoeologous read assignment.

Phylogeny and network analysis of concatenated SNPs.

Phylogenetic and network analyses were conducted using the homoeologue dataset, with SNPs concatenated to create a single supermatrix. The maximum likelihood (ML) tree, rooted with *G. max*, identified four subclades comprising two major clades: 1) the A-genome, with subclades of D4 vs. *G. clandestina* and *G. canescens*; and 2) the D-genome (D3) and E-genome (D1) (Fig. 5a). Each of the subclades showed a different pattern with respect to diploid and tetraploid subgenome relationships. In the *canescens/clandestina* clade, the A-subgenome of the synthetic allopolyploid (A58) was sister to the accession from which it was created (*G. canescens* 1232), as expected, though with deeper coalescence than expected from an artificial hybrid; the two natural T5 allopolyploids were sister to *G. clandestina* 1126, as expected from other data (e.g., Doyle et al. 2002). In the D4 clade, diploid accession 2073 was sister to all remaining accessions, a unique placement consistent with its apparently admixed nature (Fig. 3a). The polyploid subgenomes formed a paraphyletic group, with the two diploid accessions sister to the D4 subgenome of one T2 accession (1134). A similar pattern was seen in the D3 subclade, where T2 accessions formed a paraphyletic group, and all four diploid accessions formed a clade sister to T2 accession 1134. Also embedded within the T2 accessions was a clade consisting solely of T1 accessions. T1 accessions also formed a monophyletic group within the D1 clade, where natural T5 accessions and D1 accessions also formed monophyletic groups. Surprisingly, there was not a sister relationship between the D1-subgenome of synthetic allopolyploid A58 and the D1 accession from which it was formed (1316). Similar topologies were produced by neighbor-joining analysis (data not shown).

NeighborNet was used to analyze the full homoeologue dataset to identify minority patterns of relationships in the data. When rooted with *G. max*, the topology (Fig. 5b) was very similar to the ML tree (Fig. 5a), even having such features as the sister relationship of D4 2073 to other D4 accessions, and the monophyly of T1 homoeologues in both the D1 and D3 clades. There was clear evidence of character support for alternative relationships, but those relationships were minor in comparison with the major phylogenetic signal.

Gene-based phylogenetic and network analyses

Gene trees were constructed for the 27 genes (described in the Material and Methods) using several different phylogenetic and network methods. Similar topologies for trees from individual genes were obtained with BEAST and PhyML. All 27 trees showed the split between the A-genome clade and the D1/D3 clade seen in the ML tree reconstructed from concatenated SNPs (Fig. 5a). However, many individual gene trees showed unexpected groupings of one or more accessions, particularly within the A-genome clade, where several trees grouped accessions from *G. canescens* with *G. syndetika*-D4 instead of with *G. clandestina* (for example ML and BEAST trees for the gene Glyma04g39670, Supplementary Figs. 17 and 45). Relationships within the major subclades varied among the 27 gene trees. For example, nine of the 27 trees showed separate clades for *G. canescens* (plus the A58 sequence) and *G.clandestina* (e.g., Fig. 6, a and c), but in only three of them did diploid species form monophyletic groups (Supplementary Figs. 13-67). Overall, there were far more departures from expectations in the A-genome clade than in the D1/D3 clade.

There were numerous cases where alleles from diploid accessions formed monophyletic groups (e.g., 12 of 27 BEAST topologies had alleles from all four D3 accessions in a clade, often with high posterior probability). At some loci, alleles from one or more polyploids formed monophyletic clades; for example, at Glyma06g18640 (Supplementary Fig. 50), all taxa, including both homoeologous subgenomes of each polyploid, formed separate clades, with the exception of *G. clandestina*. However, this was unusual, and paraphyletic groupings of alleles were common, particularly in polyploids. For example, at 26 of 27 loci, T2-D3 alleles were not monophyletic, at least some having closer relationships to D3 or T1-D3 alleles, and in gene Glyma01g35620, T5-D1_1969 was most closely related to D1_1156 whereas T5-D1_1487 was most closely related to D1_1157 and D1_1316 (Supplementary Fig. 41). On the assumption that alleles in tetraploids all originated from diploid progenitor species, such

paraphyletic relationships suggest the input of alleles from different genotypes of diploid progenitors, due either to multiple origins or, alternatively, to continued gene flow from diploids after polyploid formation, perhaps involving unreduced gametes.

The BEAST trees, calibrated with the 5 MYA divergence of *G. max* and the perennial subgenus (Innes et al., 2008), allowed dates of allele divergence to be estimated. Among comparisons of interest are the minimum divergences between alleles from a tetraploid and alleles from its diploid progenitor (e.g., T2-D3 vs. D3) or alleles from the same progenitor in a second tetraploid (e.g., T2-D3 vs. T1-D3); the latter represent "diploid" alleles as well, under the assumption that there has been no gene flow between the two tetraploids, something that is reasonable for *G. tomentella* tetraploids (e.g., Doyle et al. 1986). Minimum distances between polyploid and diploid alleles (over)estimate the time of entry of that allele into the polyploid, which is typically assumed to be an origin of the polyploid (Doyle and Egan 2010). Minimum dates (Supplemental Table 3) were 0.31 MY for T1 (measured at the D1 locus), 0.29 MY for T5 (measured at the D1 locus), and 0.38 MY for T2 (measured at the D3 locus). Error bars on these estimates, however, were substantial.

NeighborNet (implemented in SplitsTree 4; Huson and Bryant, 2006) was used to construct networks for each of the 27 genes. Several networks showed patterns consistent with intragenic recombination; the Pairwise Homoplasy Index (PHI) of Bruen et al. (2006), also implemented in SplitsTree, was significant for 11 of the 27 genes (data not shown). The dominant patterns in NeighborNet topologies were similar to the overall pattern shown in phylogenetic analyses of the 27 genes, and thus to results for the full homoeologous SNP dataset. As with other methods, NeighborNet networks suggested multiple inputs of alleles from diploid progenitors into polyploids (e.g., gene Glyma02g11580, Fig. 6c).

Species tree reconstruction under the coalescent

Species trees were reconstructed using the coalescent approach implemented in *BEAST (Heled and Drummond, 2010), which used information contained in the individual gene trees from the 27 genes described above. The overall *BEAST tree (Fig. 7a) topology was similar to that of trees from concatenated SNPs. By definition, each of the allopolyploid homoeologous genomes was a single OTU despite the possibility of independent origins; each of these was grouped with its putative progenitor species. Within the D1 genome clade, the T1 and T5 polyploids were sisters to one another; similarly, T1 and T2 were sisters in the D3 clade. The DensiTree output (Supplementary Fig. 40) indicated considerable uncertainty only within the D3 clade, where both other possible topologies (T2 sister to D3, T1 sister to D3) appeared in a substantial number of trees. As expected, divergence dates of polyploids from their diploid progenitors estimated by *BEAST were higher than minimum estimates from the 27 individual loci, all being greater than 300,000 years (Fig. 7a).

DISCUSSION

The *Glycine* subgenus *Glycine* polyploid complex appears ideally suited as a model for studying allopolyploid evolution, because it comprises eight independently formed but closely related allopolyploid species triads (an allotetraploid and its two diploid progenitors; Fig. 1) that overlap in their genomic compositions. We are exploiting this model system to study the effect of allopolyploidy on a wide range of phenotypes, including transcriptome size, morphology, anatomy, climate niche, photosynthesis, and photoprotection (Coate and Doyle, 2010; Coate et al., 2012; Ilut et al., 2012; Coate et al., 2013; Coate and Doyle, 2013; Hegarty et al., 2013; Coate et al., 2014; Harbert et al., 2014).

To enhance the utility of this model group, it is important to move to a genome-wide understanding of their

biology. As noted above, origins of the *Glycine* allopolyploids were hypothesized initially from crossing data and more recently from gene phylogenies, but inferences have been made from only two nuclear genes. Both of these markers supported the hypotheses of fixed hybridity of *Glycine* allopolyploid species. However, it is not known to what extent the entire genomes of these plants retain contributions from both parental diploid species in the face of potential loss due to initial genomic shock (McClintock, 1984), or other processes such as "genome downsizing" (Leitch and Bennett, 2004), fractionation (Schnable and Freeling, 2011; Freeling et al., 2012), or concerted evolution (e.g., Wang et al., 2007).

Glycine allopolyploids are fixed hybrids throughout their genomes

Analyses using all SNPs identified from the full dataset showed that all three of these allopolyploids are indeed fixed hybrids, combining diploid genomes as depicted in Fig. 1. Structure results indicated an essentially equal contribution from both parental diploids in all three cases (Fig. 3a); PCA analysis also was consistent with this hypothesis, placing each polyploid approximately midway between its putative progenitors, as expected for an F1 hybrid (Fig. 4a).

In order to determine whether or not the polyploids have contributions from their parents across their entire genomes, reads were partitioned by homoeologous genome and mapped to the soybean reference genome (Schmutz et al., 2010). As portrayed by e-chromosome painting (Fig. 2), it is clear that no individual sampled from any of the three allopolyploid species has any major regions represented by only one homoeologue. Coverage is sparse in pericentromeric and centromeric regions, as expected due to the low density of genes in these regions of the soybean genome (Schmutz et al., 2010). The degree of shared synteny between soybean and these perennial *Glycine* species is as yet unknown, but regardless of the order of chromosomal segments, it is clear that there has not been significant loss of homoeologous genes. We mapped reads to over 22,000 of the approximately 46,000 genes of the soybean genome (Schmutz et al., 2010). These numbers include both homoeologous copies from the 5-10 MYA polyploidy event that shaped the modern "diploid" ($2n = 38,40$) *Glycine* genome. We were able to deconvolute between 4 and 19% of these 22,000 genes into their homoeologous contributions in each of the three recent allopolyploids (e.g., T1D1 and T1D3). Using genomic in situ hybridization (GISH), Chester et al. (2012) showed examples of allopolyploid *T. miscellus* plants that had all four chromosomes or chromosome segments of one diploid parent (4:0), but also examples of plants with 3:1 ratios of homoeologous chromosomes or chromosomal segments. Our e-chromosome painting method cannot distinguish the 3:1 condition from an equal contribution from both parents segments, so it is possible that such plants exist in our sample.

Structure analysis using the partitioned homoeologous SNPs corroborated results with the full, unpartitioned dataset, in placing each polyploid homoeologous genome with its putative progenitor (Fig. 3b). The FineStructure PCA supported three major groupings, each of which included diploids and the expected polyploid homoeologous subgenomes derived from them (Fig. 4b). The grouping of D4 accessions and two A-genome species (*G. canescens* and *G. clandestina*), along with polyploid genomes derived from them, into a single cluster is not surprising, because *G. syndetika* (D4) is also a member of the A-genome (Ratnaparkhe et al., 2011). As noted above, genome groups were originally defined on the basis of reproductive compatibility in artificial crosses (Ratnaparkhe et al. 2011), and indeed *G. syndetika* (D4) 2073 shows evidence of admixture with *G. canescens* and *G. clandestina* (Fig. 3). In contrast, D1 and D3, though both classified as "*G. tomentella*", belong to two different genome groups (E and D, respectively; Ratnaparkhe et al. 2011). This greater genetic similarity of the three A-genome species is not reflected in relative divergence dates; for example, the *BEAST analysis dates the divergence between *G. syndetika* and the two other A-genome species at slightly earlier than the divergence between D1 and D3 (Fig. 7a). Thus, reproductive barriers likely arose earlier in the D1/D3 lineage than within the A-genome.

Allopolyploid evolution in *Glycine* fits "Darlington's Rule" (Darlington, 1937)—that allopolyploids should form between species that are reproductively isolated, often due to chromosomal differences, whereas reproductively

compatible diploids instead tend to form homoploid hybrids. No allopolyploids are known to have formed among A-genome species, and only one of the eight known *Glycine* allopolyploids involves hybridization within a genome group (tetraploid *G. tabacina* is the product of the most divergent species cross possible within the B-genome; Doyle et al., 2004). D1 and D3, which as noted belong to different genome groups, have different chromosome numbers ($2n = 38$ vs. 40, respectively), which may contribute to their inability to form fertile diploid hybrids. D1 has also formed allopolyploids with D5A, another $2n = 40$ "*G. tomentella*"; however, reproductive incompatibility also occurs between $2n = 40$ *G. tomentella* taxa (Doyle et al., 1986), and other allopolyploids in the complex combine genomes of two $2n = 40$ taxa (Fig. 1).

Gene histories, allele divergence times, and sources of genetic diversity in polyploids.

Gene trees from the 27 loci were selected that met criteria designed to provide orthologues. These genes are highly transcribed with sufficient characters for phylogeny reconstruction, and inferences of polyploid origins mostly conformed to expectations based on previous work using the low copy nuclear locus, histone H3D (Brown et al., 2002; Doyle et al., 2002; González-Orozco et al., 2012), the nrDNA ITS (Singh et al., 2001; Rauscher et al., 2004), and chloroplast noncoding sequences (Hsing et al. 2001). The use of BEAST and *BEAST (Heled and Drummond, 2010) allowed us to estimate divergence times of alleles and species for the first time for some of these taxa. Dating polyploid origins is complicated by numerous factors (Doyle and Egan, 2010). For one thing, if the polyploid has arisen recurrently, then there is no single date that marks "the" origin. Given that polyploids are often invasive (e.g., Pandit et al., 2011), and the *Glycine tomentella* allopolyploids appear to be recently formed based on sharing identical histone H3D and nrDNA ITS alleles with their putative progenitors (Doyle et al., 2002; Rauscher et al., 2004), we have speculated that they could have originated as a response to ecological disturbance due to human colonization of Australia, around 40,000 years ago (Hudjashov et al., 2007; Pugach et al., 2013). The relevant date for testing this anthropogenic disturbance hypothesis would be the oldest origin of each polyploid. However, because it is unlikely that a polyploid allele and any of a set of diploid progenitor alleles will coalesce at exactly the time of polyploid origin, distances for any given polyploid event will be overestimates of the actual time of origin. Further complicating matters, the error bars on our BEAST divergence estimates were large relative to the estimates themselves. Nevertheless, because even the minimum estimates of allele divergences between diploids and tetraploids are around 0.3 MY, it appears likely that these *G. tomentella* allopolyploids are hundreds of thousands rather than tens of thousands of years old. *BEAST estimates should be averages of all origins of a polyploid taxon, and these, too are several hundred thousand years for each allopolyploid. Thus, it appears likely that these polyploid species were present in Australia before humans arrived there. The fact that these three species, and possibly other allopolyploid members of the complex, may have evolved at roughly the same time is intriguing. In the ca. 5 MY since the perennial members of *Glycine* diverged from the annual lineage (Egan and Doyle, 2010), there is no evidence of polyploidy until these species were formed, apparently well within the last 1 MY. Perhaps the onset of severe aridity in Australia around 3 MYA, heralding the change to the present extreme wet-dry glacial cycles (Crisp et al., 2004) could have provided ecological opportunities for polyploids. It will be interesting to refine our estimates through increased sampling of these three triads, and to obtain estimates for the other five allopolyploid species.

ACKNOWLEDGEMENTS

The authors thank Sue Sherman-Broyles for helpful comments throughout the project. We also are grateful to Steven Cannon and a second, anonymous reviewer for detailed comments and suggestions.

REFERENCES

Arnold B, Bomblies K, Wakeley J (2012) Extending Coalescent Theory to Autotetraploids. Genetics **192**: 195–204

Aronesty E (2013) Comparison of sequencing utility programs. Open Bioinform J. **7**:1-8

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

Bombarely A, Edwards KD, Sanchez-Tamburrino J, Mueller LA (2012) Deciphering the complex leaf transcriptome of the allotetraploid species Nicotiana tabacum: a phylogenomic perspective. BMC genomics **13**:406

Brown AHD, Doyle JL, Grace JP, Doyle JJ (2002) Molecular phylogenetic relationships within and among diploid races of Glycine tomentella (Leguminosae). Australian Systematic Botany **15**: 37–47

Cannon SB, Ilut D, Farmer AD, Maki SL, May GD, Singer SR, Doyle JJ (2010) Polyploidy did not predate the evolution of nodulation in all legumes. PLoS ONE **5**: e11630

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE (2012) Extensive chromosomal variation in a recently formed natural allopolyploid species, Tragopogon miscellus (Asteraceae). Proc. Natl. Acad. Sci. U.S.A. **109**: 1176–1181

Coate JE, Doyle JJ (2010) Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. Genome Biol Evol **2**: 534–546

Coate JE, Doyle JJ (2013) Genomics and transcriptomics of photosynthesis in polyploids. In Polyploid and Hybrid Genomics, (Chen ZJ and Birchler JA, eds) Hoboken, NJ: Wiley-Blackwell, pp. 153–169

Coate JE, Luciano AK, Seralathan V, Minchew KJ, Owens TG, Doyle JJ (2012) Anatomical, biochemical, and photosynthetic responses to recent allopolyploidy in Glycine dolichocarpa (Fabaceae). Am J Bot **99**: 55–67

Coate JE, Powell AF, Owens TG, Doyle JJ (2013) Transgressive physiological and transcriptomic responses to light stress in allopolyploid Glycine dolichocarpa (Leguminosae). Heredity **110**: 160–170

Coate JE, Bar H, Doyle JJ (2014) Extensive translational regulation of gene expression in an allopolyploid correlates with long term retention of duplicated genes. The Plant Cell **26**:136:150

Crisp M, Cook L, Steane D (2004) Radiation of the Australian flora: what can comparisons of molecular phylogenies across multiple taxa tell us about the evolution of diversity in present-day communities? Philos Trans R Soc Lond, B, Biol Sci **359**: 1551–1571

Darlington CD (1937) Recent advances in cytology. Second edition. Philadelphia. P. Blakiston's son and Co.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Meth **9**: 772

Doyle JJ, Schuler MA, Godette WD, Zenger V, Beachy RN, Slightom JL (1986) The glycosylated seed storage proteins of Glycine max and Phaseolus vulgaris. Structural homologies of genes and proteins. Journal of Biological Chemistry **261**(20):9228-38.

Doyle JJ, Doyle JL, Brown AH (1999) Origins, colonization, and lineage recombination in a widespread perennial soybean polyploid complex. Proc. Natl. Acad. Sci. U.S.A. **96:** 10741–10745

Doyle JJ, Doyle JL, Brown A, Palmer RG (2002) Genomes, Multiple Origins, and Lineage Recombination in the

503 Glycine Tomentella (Leguminosae) Polyploid Complex: Histone H3-D Gene Sequences. Evolution

504 **Doyle JJ, Doyle JL, Rauscher J, Brown A** (2004) Diploid and Polyploid Reticulate Evolution Throughout the
505 History of the Perennial Soybeans (Glycine Subgenus Glycine). New Phytol **161**: 121–132

506 **Doyle JJ, Egan AN** (2010) Dating the origins of polyploidy events. New Phytol **186**: 73–85

507 **Doyle JJ** (2012) Polyploidy in Legumes. *In* PS Soltis, DE Soltis, eds, Polyploidy and Genome Evolution. Springer
508 Berlin Heidelberg, Berlin, Heidelberg, pp 147–180

509 **Doyle MJ, Brown AHD** (1985) Numerical analysis of isozyme variation in Glycine tomentella. Biochem Syst Ecol
510 **13**: 413–419

511 **Doyle MJ, Grant JE, Brown AHD** (1986) Reproductive isolation between isozyme groups of Glycine tomentella
512 (Leguminosae) and spontaneous doubling in their hybrids. Aust J Bot **34**:523–535

513 **Drummond AJ, Suchard MA, Xie D, Rambaut A** (2012) Bayesian phylogenetics with BEAUti and the BEAST
514 1.7. Mol Biol Evol **29**: 1969–1973

515 **Du J, Tian Z, Sui Y, Zhao M, Song Q, Cannon SB, Cregan P, Ma J** (2012) Pericentromeric effects shape the
516 patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean. Plant Cell **24**:
517 21–32

518 **Dufresne F, Stift M, Vergilino R, Mable BK** (2014) Recent progress and challenges in population genetics of
519 polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol Ecol **23**: 40–69

520 **Edwards SV** (2009) Is a new and general theory of molecular systematics emerging? Evolution **63**: 1–19

521 **Egan AN, Doyle JJ** (2010) A comparison of global, gene-specific, and relaxed clock methods in a comparative
522 genomics framework: dating the polyploid history of soybean (Glycine max). Syst Biol **59**: 534–547

523 **Evanno G, Regnaut S, Goudet J** (2005) Detecting the number of clusters of individuals using the software
524 STRUCTURE: a simulation study. Mol Ecol **14**: 2611–2620

525 **Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC** (2012) Fractionation mutagenesis
526 and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr Opin Plant
527 Biol **15**: 131–139

528 **Gilad Y, Pritchard JK, Thornton K** (2009) Characterizing natural variation using next-generation sequencing
529 technologies. Trends Genet **25**: 463–471

530 **González-Orozco CE, Brown AHD, Knerr N, Miller JT, Doyle JJ** (2012) Hotspots of diversity of wild Australian
531 soybean relatives and their conservation in situ. Conserv Genet **13**: 1269–1281

532 **Grant JE, Brown AHD, Grace JP** (1984) Cytological and isozyme diversity in Glycine tomentella Hayata
533 (Leguminosae). Aust J Bot **32**:665–677

534 **Grover CE, Salmon A, Wendel JF** (2012) Targeted sequence capture as a powerful tool for evolutionary analysis.
535 Am J Bot **99**: 312–319

536 **Guindon S, Gascuel O** (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum
537 likelihood. Syst Biol **52**: 696–704

538 **Harbert RS, Brown AHD, Doyle JJ** (2014) Allopolyploidy, climate niche modeling, and evolutionary "success" in
539 *Glycine* (Leguminosae). Am J Bot **101**: 710–721

540 **Hegarty M, Coate J, Sherman-Broyles S, Abbott R, Hiscock S, Doyle J** (2013) Lessons from natural and artificial
541 polyploids in higher plants. Cytogenet Genome Res **140**: 204–225

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

Hollister JD, Arnold BJ, Svedin E, Xue KS, Dilkes BP, Bomblies K (2012) Genetic adaptation associated with genome-doubling in autotetraploid Arabidopsis arenosa. PLoS Genet **8**: e1003093

Hsing Y-LC, Hsieh J-S, Peng C-L, Chou C-H, Chiang T-Y (2001) Systematic Status of the Glycine tomentella and G. tabacina Species Complexes (Fabaceae) Based on ITS Sequences of Nuclear Ribosomal DNA. J Plant Res **114**: 435–442

Hudjashov G, Kivisild T, Underhill PA, Endicott P, Sanchez JJ, Lin AA, Shen P, Oefner P, Renfrew C, Villems R, et al (2007) Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. Proc. Natl. Acad. Sci. U.S.A. **104**: 8726–8730

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol **23**: 254–267

Hymowitz T, Singh RJ, Kollipara KP (2010) The Genomes of the Glycine. *In* Plant Breeding Reviews. John Wiley & Sons, Inc, Oxford, UK, pp 289–317

Ilut DC, Coate JE, Luciano AK, Owens TG, May GD, Farmer A, Doyle JJ (2012) A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. Am J Bot **99**: 383–396

Innes RW, Ameline-Torregrosa C, Ashfield T, Cannon E, Cannon SB, Chacko B, Chen NWG, Couloux A, Dalwani A, Denny R, et al (2008) Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. Plant Physiol **148**: 1740–1759

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al (2012) A genome triplication associated with early diversification of the core eudicots. Genome Biol **13**: R3

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature **473**: 97–U113

Joly S, Rauscher JT, Sherman-Broyles SL, Brown AHD, Doyle JJ (2004) Evolutionary dynamics and preferential expression of homeologous 18S-5.8S-26S nuclear ribosomal genes in natural and artificial glycine allopolyploids. Mol Biol Evol **21**: 1409–1421

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

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Meth **9**: 357–359

Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. PLoS Genet **8**: e1002453

Leitch IJ, Bennett MD (2004) Genome downsizing in polyploid plants. Biological Journal of the Linnean Society **82**:651-663

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science **290**: 1151–1155

582 **Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y** (2005) Modeling gene and
583 genome duplications in eukaryotes. P Natl Acad Sci Usa **102**: 5454–5459

584 **McClintock B** (1984) The significance of responses of the genome to challenge. Science **226**: 792–801

585 **Pandit MK, Pocock MJO, Kunin WE**, (2011) Ploidy influences rarity and invasiveness in plants. J. Ecol. **99**:1108-
586 1115.

587 **Pritchard J, Stephens M, Donnelly P** (2000) Inference of population structure using multilocus genotype data.
588 Genetics **155**: 945–959

589 **Pugach I, Delfin F, Gunnarsdottir E, Kayser M, Stoneking M** (2013) Genome-wide data substantiate Holocene
590 gene flow from India to Australia. Proc. Natl. Acad. Sci. U.S.A. **110**:1803-1808.

591 **Rambaut A** (2012) FigTree version 1.4.0. http://tree.bio.ed.ac.uk/software/figtree/

592 **Ramsey J, Schemske DW** (2002) Neopolyploidy in flowering plants. Annu. Rev. Ecol. Syst. **33**:589-639.

593 **Ratnaparkhe MB, Singh RJ, Doyle JJ** (2011) Glycine. Wild Crop Relatives: Genomic and Breeding Resources,
594 Legume Crops and Forage, C. Kole (ed.) Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, pp 83–116

595 **Rauscher JT, Doyle JJ, Brown AHD** (2004) Multiple origins and nrDNA internal transcribed spacer homeologue
596 evolution in the Glycine tomentella (Leguminosae) allopolyploid complex. Genetics **166**: 987–998

597 **Rauscher JT, Doyle JJ, Brown AHD** (2002) Internal transcribed spacer repeat-specific primers and the analysis of
598 hybridization in the Glycine tomentella (Leguminosae) polyploid complex. Mol Ecol **11**: 2691–2702

599 **Schlueter J, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker R** (2004) Mining EST databases to
600 resolve evolutionary events in major crop species. Genome **47**: 868–876

601 **Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al**
602 (2010) Genome sequence of the palaeopolyploid soybean. Nature **463**: 178–183

603 **Schnable JC, Freeling M** (2011) Genes identified by visible mutant phenotypes show increased bias toward one of
604 two subgenomes of maize. PLoS ONE **6**: e17855

605 **Shoemaker RC, Schlueter J, Doyle JJ** (2006) Paleopolyploidy and gene duplication in soybean and other legumes.
606 Curr Opin Plant Biol **9**: 104–109

607 **Singh RJ, Kim HH, Hymowitz T** (2001) Distribution of rDNA loci in the genus Glycine Willd. Theor Appl Genet
608 **103**: 212–218

609 **Singh RJ, Kollipara KP, Hymowitz T** (1998) The genomes of Glycine canescens FJ Herm., and G. tomentella
610 Hayata of Western Australia and their phylogenetic relationships in the genus Glycine Willd. Genome **41**:669-679

611 **Slotte T, Bataillon T, Hansen TT, St Onge K, Wright SI, Schierup MH** (2011) Genomic determinants of protein
612 evolution and polymorphism in Arabidopsis. Genome Biol Evol **3**: 1210–1219

613 **Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman**
614 **LK, et al** (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat
615 Genet **45**: 831–835

616 **Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall**
617 **PK, Soltis PS** (2009) Polyploidy and angiosperm diversification. Am J Bot **96**: 336–348

618 **Soltis DE, Buggs RJA, Doyle JJ, Soltis PS** (2010). What we still don't know about polyploidy. Taxon **59**:1387-
619 1403.

620 **Symonds VV, Soltis PS, Soltis DE** (2010) Dynamics of polyploid formation in Tragopogon (Asteraceae): recurrent
621 formation, gene flow, and population structure. Evolution **64**: 1984–2003

622 **Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L**
623 (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat
624 Protoc **7**: 562–578

625 **Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L**
626 (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching
627 during cell differentiation. Nat Biotechnol **28**: 511–174

628 **Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH** (2007) Extensive concerted evolution of rice paralogs and
629 the road to regaining independence. Genetics **177**: 1753–1763

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

Figure 1

Schema of the *Glycine* perennial polyploid complex.

Diploid progenitors are represented by circles and allotetraploid species by squares. Chromosome numbers are shown for each species, and genome groups (Ratnaparkhe et al., 2011) are given for diploids. Species used in this study (*G. tomentella* D1, *G. tomentella* D3, *G. syndetika* D4, *G. canescens*, *G. clandestina*, *G. dolichocarpa* T2, *G. tomentella* T1 and *G. tomentella* T5) are shown in green.

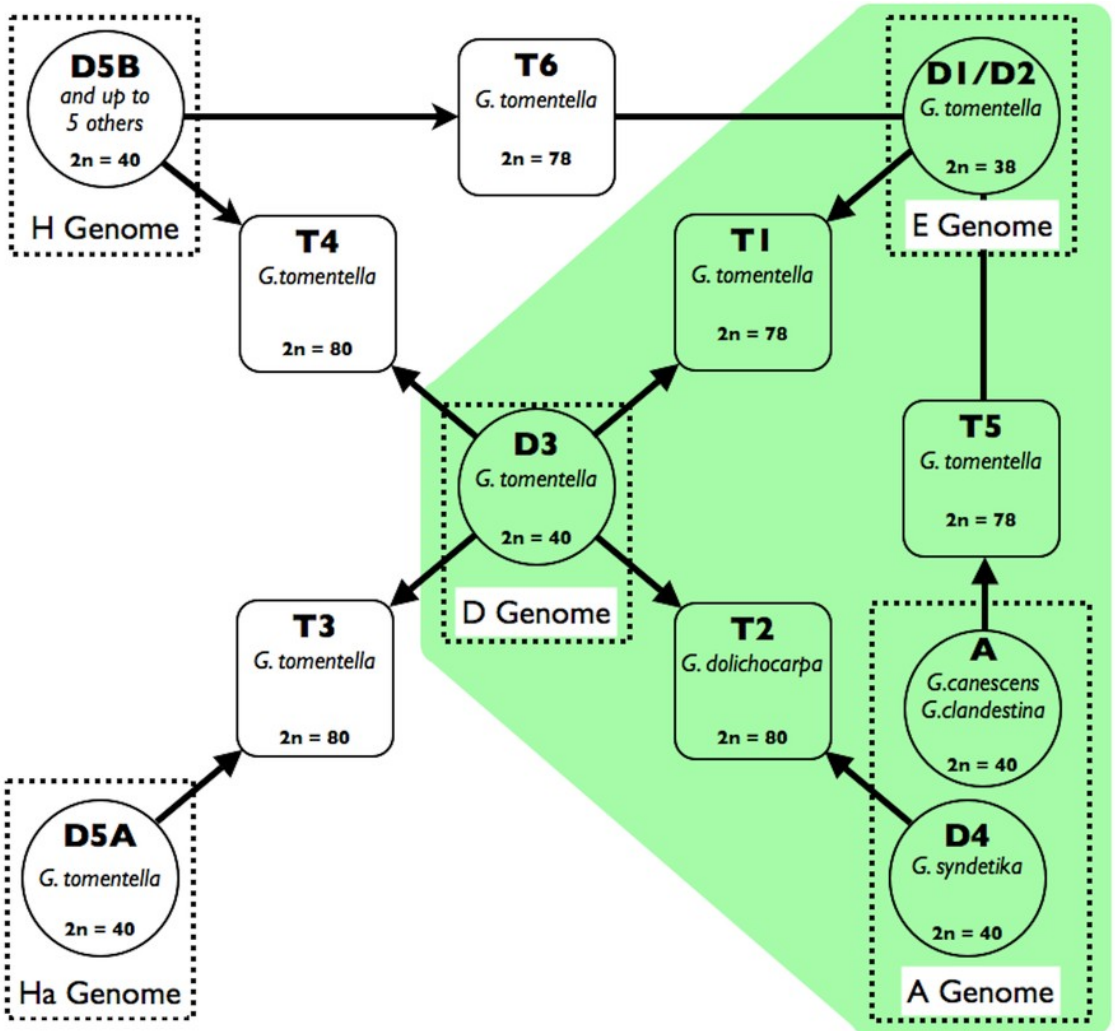

Figure 2

Electronic chromosome painting for *G. dolichocarpa* T2 accession 1134.

SNP positions on the 20 soybean chromosomes are represented by blue lines (D3 progenitor) or red lines (D4 progenitor).

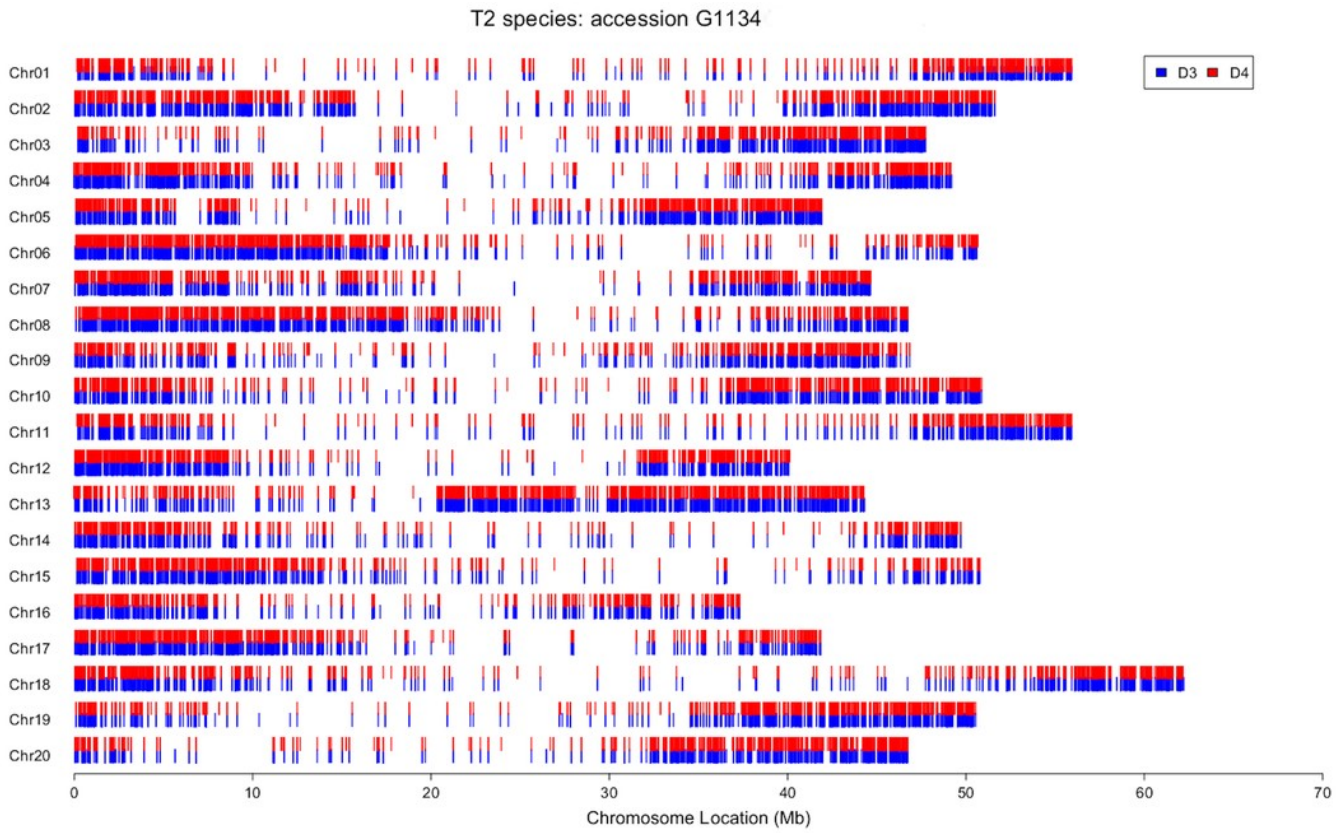

Figure 3

Structure analysis for *Glycine* perennial polyploid accessions

SNP analysis using Structure for a set of 20,000 random SNPs for *Glycine* polyploid complex accessions (A) without homoeologue separation and (B) with homoeologue separation for K = 6. The five progenitor diploid species are placed in different populations: red (*G. clandestina*), dark red (*G. canescens*), yellow (*G. tomentella D1*), blue (*G. syndetika D4*) and green (*G. tomentella D3*).

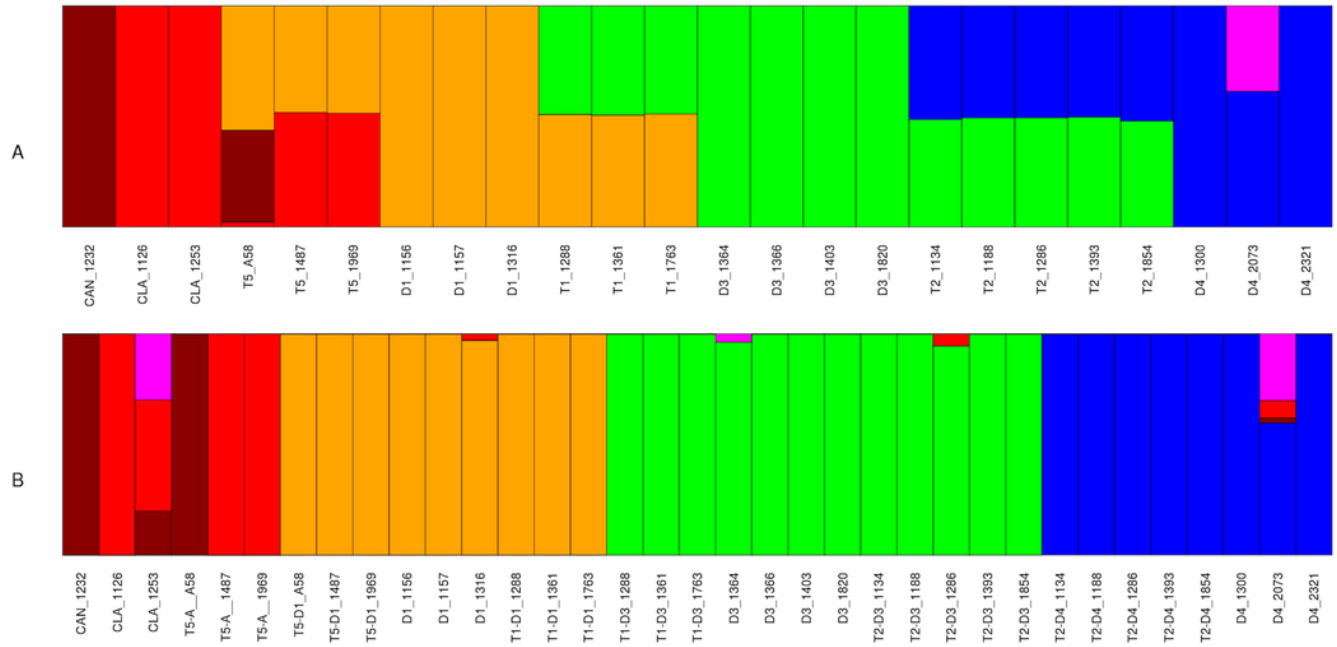

Figure 4

FineStructure Analysis for *Glycine* perennial polyploid species

Analysis using FineStructure for (A, C) 220,952 SNPs for the *Glycine* perennial polyploid complex (species groups A, D1, D3, D4, T1, T2 and T5) without homoeologue separation. 7 clusters can be distinguished (one per species group) in the PCA analysis where polyploids are admixtures of the diploid progenitor groups (A). The heatmap (C) shows diploid hybrid signal for polyploids, for example T5_A58 shows a stronger signal with its progenitors: CAN_1232 (blue) and D1_1316 (intense orange). (B, D) 70,910 SNPs for the *Glycine* perennial polyploid complex after homoeologue separation. 3 clusters can be distinguished in the PCA analysis (B): right cluster, species from the A-genome (A and D4); bottom-left cluster, species D1; and top-left cluster, species D3. (D) The heatmap signal is divided into the same three major clusters.

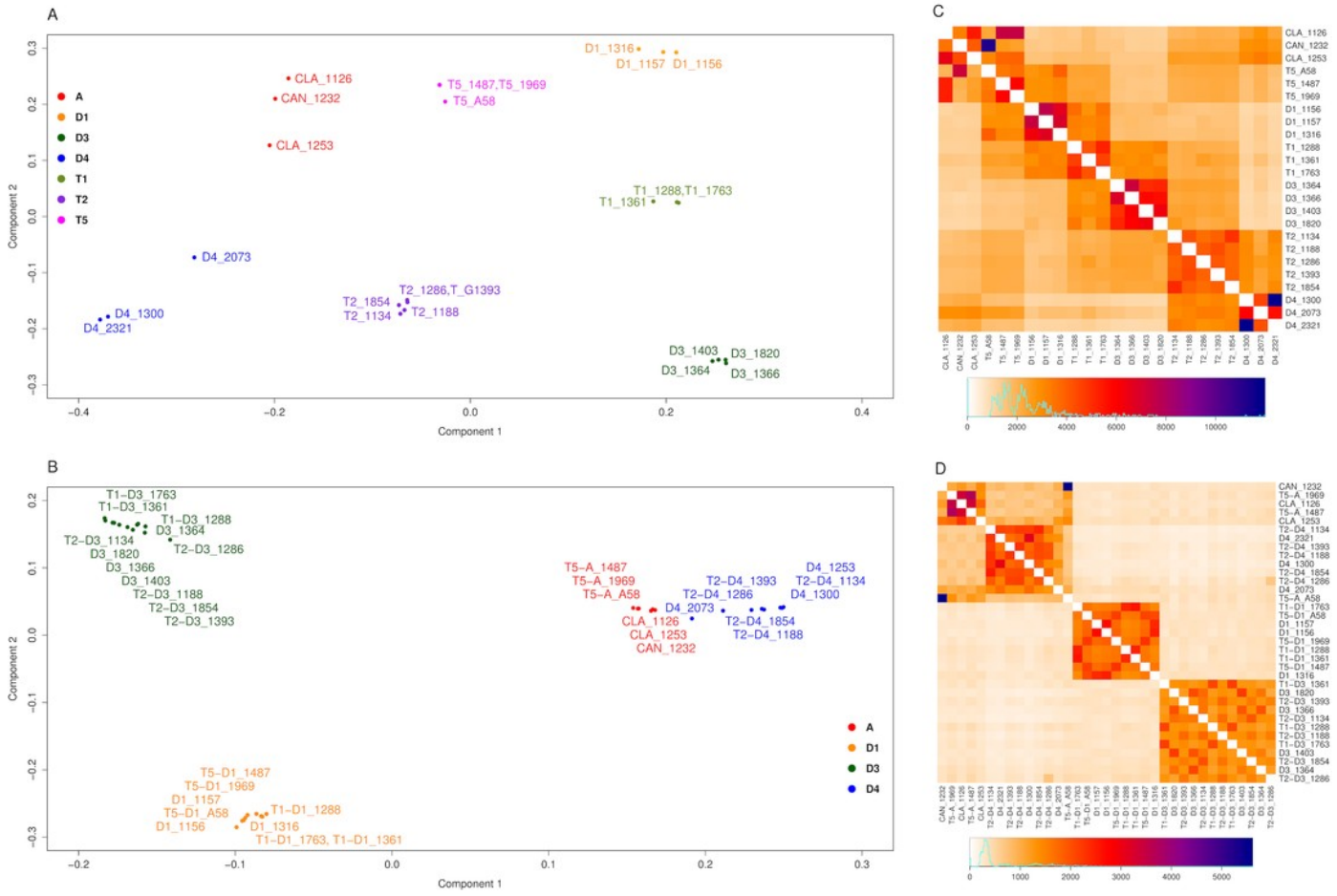

Figure 5

Phylogenetic relationship in the *Glycine* perennial polyploid complex

Relationships in the Glycine perennial polyploid complex after homoeologue separation, using a concatenated dataset. Branches are colored as in Fig. 4, based on the 5 different diploid species. In both the maximum likelihood (ML) phylogeny (A) and the NeighborNet network (B), the same major species groups are visible (D1, D3, D4 and *G. canescens/G. clandestina*).

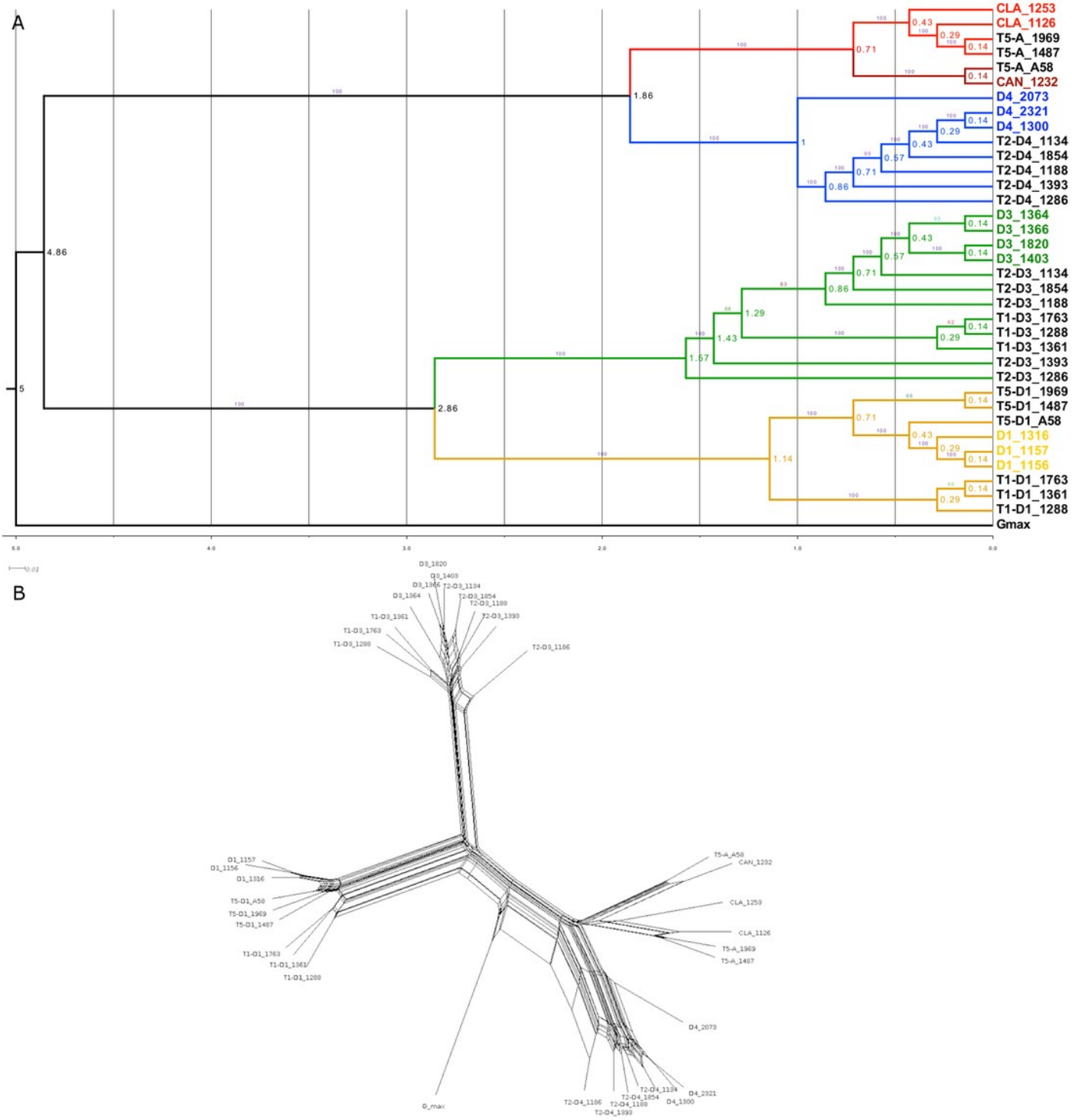

Figure 6

Phylogenetic analysis for the Glyma02g11580 locus

Glyma02g11580 locus using ML with bootstrap values (A), NeighborNet (B), and BEAST with posterior probabilities and showing node ages (in black) (C). For figures A and C, branch length colors represent bootstrap or posterior probabilities values, with red shades being the lowest values and green shades being the highest values.

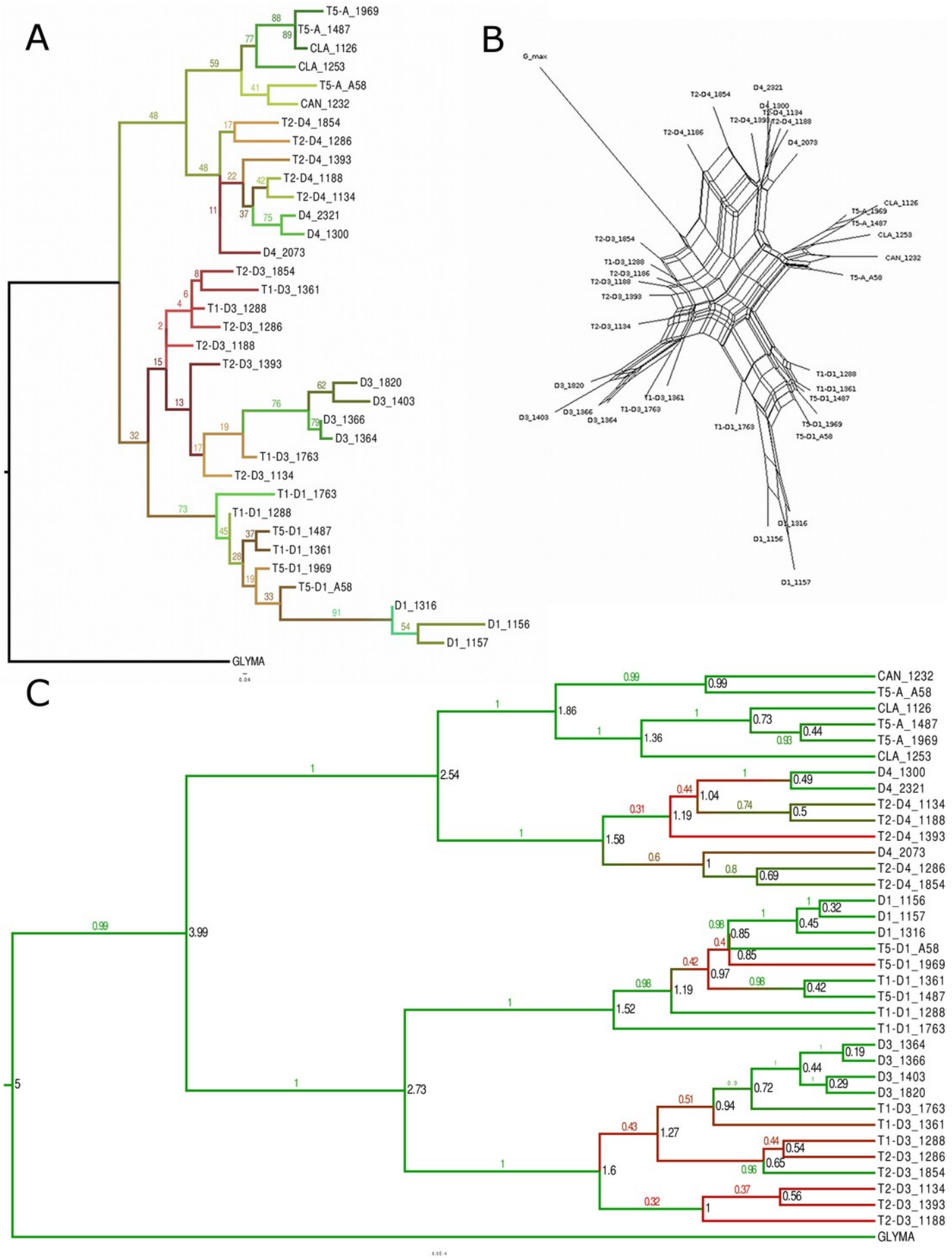

Figure 7

Phylogenetic tree with estimated divergence dates

*BEAST tree with the estimated node ages and error bars representing the highest posterior density (HPD) interval at the 95% level.

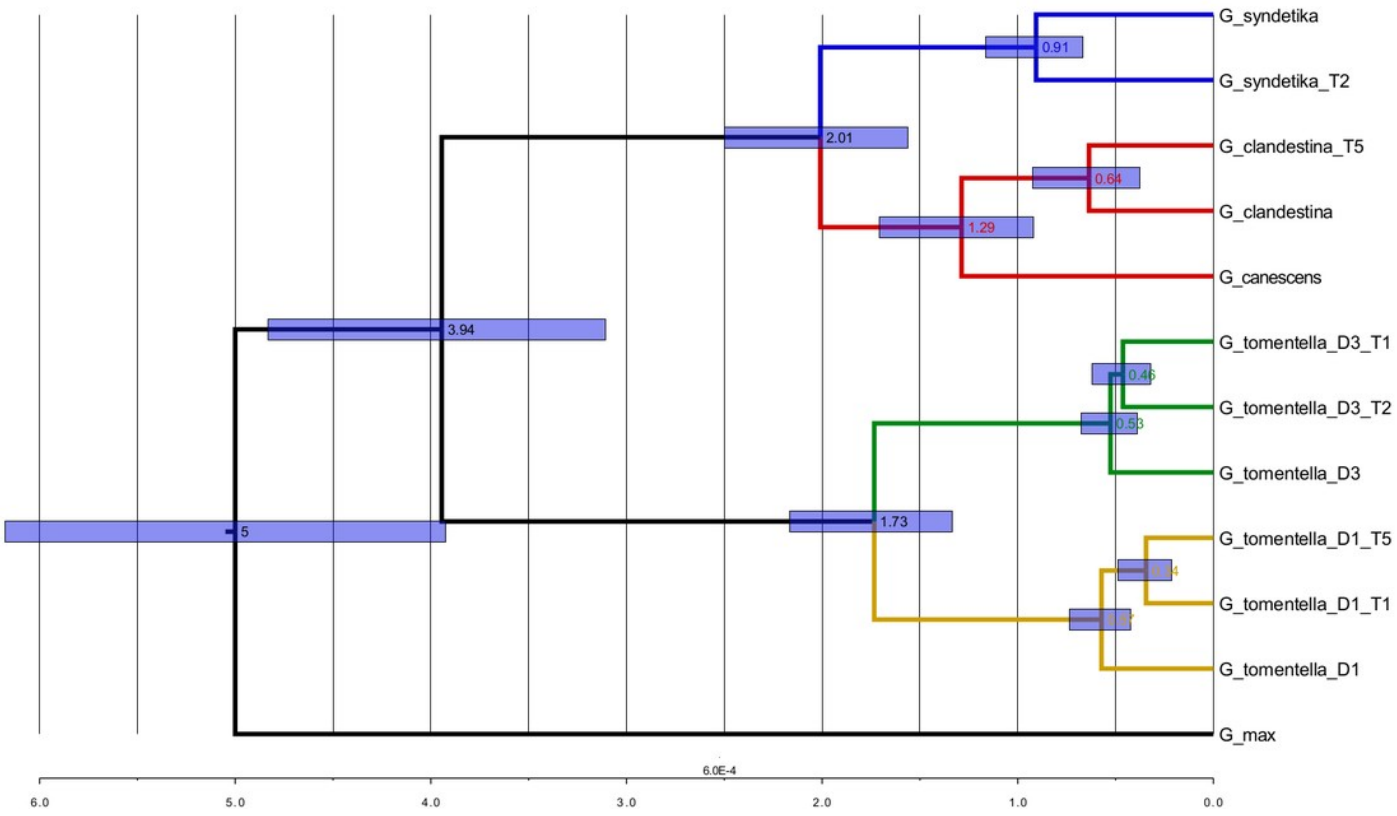

Table 1(on next page)

Sequencing, reads processing and mapping summary.

Represented genes reflected the number of *Glycine max* reference genome genes where after the perennials reads mapping and expression measure have an expression > 0 (RPKM). Gray shading = allopolyploid species.

Species	Accession	Samples	Raw Reads	Processed Reads	Mapped Reads	Represented Genes
Glycine canescens	1232	2	21332880	20696801	14381555	23833
Glycine clandestina	1126	2	19086864	18613018	11815996	23340
	1253	3	33546015	32326942	19117095	23723
Glycine dolichocarpa	1134	13	202427873	187120918	60712525	23643
	1188	2	19034633	18279858	11960713	22952
	1286	2	11814995	11216980	7422888	25278
	1393	2	21820163	21029983	13643602	23345
	1854	3	54748079	42826840	16032643	22718
Glycine syndetika	1300	3	25527322	23634740	14092961	24238
	2073	2	12132989	11072073	7087710	24438
	2321	2	32796391	30024544	13637368	22571
Glycine tomentella D1	1156	3	38218179	36988846	21905536	23041
	1157	2	16522541	15906072	9715890	23920
	1316	2	25207045	24375482	15417078	22749
Glycine tomentella D3	1364	1	10401944	9604350	6896983	22802
	1366	2	20631583	18098232	10766169	23364
	1403	3	31631369	28953234	17218424	23352
	1820	3	71185274	63055644	18625439	22871
Glycine tomentella T1	1288	2	14608219	14148847	9298349	23348
	1361	2	17964870	17627119	11217736	23758
	1763	2	21870236	20933661	14101838	23349
Glycine tomentella T5	A58_1	2	22447334	21996303	13389955	23042
	1487	2	21267274	20469069	13907305	23437
	1969	3	21324229	20847883	11136293	23522

Table 2(on next page)

Summary of SNPs using *G. max* as reference genome.

Gray shading = allopolyploid species. (* Between square brackets the coverage of the G. max transcriptome, including alternative splicings)(** Square brackets = percentage of heterozygous positions).

Species	Accession	% Gmax Coverage*	Raw SNPs	Processed SNPs**	Synonymous	Non-Synonymous
Glycine canescens	1232	7.2 [65.0]	589686	453,398 [7.7]	148321	123413
Glycine clandestina	1126	6.7 [61.4]	496746	375,943 [7.5]	115340	96562
	1253	7.5 [65.1]	617543	487,923 [8.3]	143952	124920
Glycine dolichocarpa	1134	11.6 [77.4]	1135676	965,643 [26.4]	242556	221326
	1188	7.4 [65.1]	550698	423,353 [28.9]	132471	113785
	1286	4.5 [45.5]	302661	224,653 [27.9]	67187	53595
	1393	6.7 [62.9]	470402	367,646 [28.8]	125339	104549
	1854	7.8 [65.3]	580531	471,020 [25.8]	140911	120274
Glycine syndetika	1300	7.5 [65.5]	605556	477,245 [7.8]	147362	125041
	2073	6.0 [57.6]	402798	282,215 [12.6]	91451	75333
	2321	8.0 [67.7]	670121	544,101 [6.3]	166409	143612
Glycine tomentella D1	1156	8.6 [69.7]	767614	621,043 [7.6]	190778	160781
	1157	6.2 [56.8]	455265	328,574 [7.1]	94377	77945
	1316	7.2 [62.3]	537439	412,518 [9.3]	120056	99666
Glycine tomentella D3	1364	5.0 [51.8]	335301	226,697 [7.8]	84917	65888
	1366	6.6 [59.7]	481258	360,327 [7.5]	111011	90015
	1403	6.4 [60.8]	476495	369,661 [6.6]	121526	99074
	1820	9.3 [69.6]	803774	641,145 [6.6]	188965	161826
Glycine tomentella T1	1288	6.9 [63.0]	498900	371,845 [19.6]	121418	102548
	1361	5.1 [54.2]	293339	200,738 [18.4]	75653	59378
	1763	7.1 [65.5]	533041	417,420 [19.4]	140203	116465
Glycine tomentella T5	A58_1	7.3 [64.6]	544331	430,552 [27.3]	135163	113781
	1487	7.0 [63.6]	516755	395,503 [26.8]	128199	105647
	1969	7.4 [65.9]	558920	444,468 [27.5]	146711	124933

Table 3(on next page)

Summary of the mapped reads and SNPs produced after the homoeologus reads separation.

It is based in the selective mapping with its progenitors (* Square brackets = percentage of heterozygous positions).

Species	Accession	Progenitor I	Mapped to Progenitor I (%)	SNPs for I*	Progenitor II	Mapped to Progenitor II (%)	SNPs for II*
Glycine dolichocarpa	1134	D3	11.4	399,884 [2.2]	D4	11.6	380,389 [2.1]
	1188	D3	20.8	227,610 [2.0]	D4	20.4	220,610 [2.1]
	1286	D3	20.3	124,984 [1.7]	D4	20.3	123,873 [1.8]
	1393	D3	19.6	197,132 [1.9]	D4	19.8	192,148 [1.9]
	1854	D3	17.9	245354 [1.5]	D4	19.3	242,561 [1.7]
Glycine tomentella T1	1288	D1	14.9	143,232 [1.7]	D3	17.5	160,873 [1.9]
	1361	D1	15.0	155,360 [1.6]	D3	17.6	175,871 [2.0]
	1763	D1	14.8	158,777 [1.8]	D3	17.3	179,032 [2.0]
Glycine tomentella T5	A58_1	A	16.9	190,138 [2.1]	D1	20.5	222,134 [1.8]
	1487	A	17.1	174,051 [1.9]	D1	20.0	202,555 [1.7]
	1969	A	16.0	182,615 [2.4]	D1	18.6	214,799 [1.8]

Table 4(on next page)

Summary of SNP count between species groups.

Polyploids are divided in two species according the progenitor origin. A Species includes *G. canescens, G. clandestina* and *G. tomentella* T5-A; D1 species includes *G. tomentella* D1, *G. tomentella* T1-D1 and *G. tomentella* T5-D1; D3 species includes *G. tomentella* D3, *G. tomentella* T1-D3 and *G. tomentella* T2-D3; D4 species includes *G. syndetika* and *G. tomentella* T2-D4. * The same species group contains the specific SNPs between accession of the same species.

Species Group	*Gmax* SNPs	A Group SNPs	D1 Group SNPs	D3 Group SNPs	D4 Group SNPs
A Species	9406	26,438 *	7096	6591	1465
D1 Species	11187	-	21,830 *	5933	7556
D3 Species	9299	-	-	25,157 *	7295
D4 Species	9314	-	-	-	23,324 *

Table 5(on next page)

Summary of the genes used in the BEAST and *BEAST analysis.

Tree likelihood values and the functional annotation are shown.

GeneID	TreeLikelihood Mean	TreeLikelihood ESS	Gene Functional Annotation
Glyma01g35620	-4676.067	1361.926	Phytoene dehydrogenase
Glyma02g11580	-3986.812	1034.414	RNA binding protein
Glyma03g29330	-7611.686	894.813	Magnesium chelatase
Glyma03g36630	-2666.725	696.197	Rho GTPase activating protein
Glyma04g39670	-4142.157	2556.251	ATP-binding transport protein-related
Glyma05g05750	-3028.809	541.483	Beta-amylase
Glyma05g09310	-2578.198	305.191	Pyruvate kinase
Glyma05g26230	-3741.498	5156.337	Metalloprotease M41 FtsH
Glyma05g37840	-2138.404	3766.767	Haloacid dehalogenase-like hydrolase
Glyma06g18640	-3418.91	6613.752	Elongation factor Tu
Glyma07g03370	-2091.845	742.796	Palmytoil-monogalactosyldiacylglycerol delta-7 desaturase
Glyma07g17180	-2218.934	2833.162	Fructose-1,6-bisphosphatase
Glyma10g42100	-2903.849	1774.192	3-ketoacyl-CoA synthase
Glyma11g13880	-4644.419	7487.252	Lipoxygenase
Glyma11g33720	-3592.689	2273.389	DELLA protein
Glyma12g04150	-2061.527	4777.335	Fructose-bisphosphate aldolase
Glyma12g12230	-2177.72	1790.185	O-methyltransferase
Glyma13g17820	-2439.715	342.999	Polyubiquitin
Glyma14g03500	-2063.541	819.216	Phytoene synthase
Glyma16g00410	-4041.294	4475.536	heat shock protein 70
Glyma16g01980	-4985.489	387.993	Myb-like protein
Glyma16g04940	-2152.355	4586.42	Glyceraldehyde 3-phosphate dehydrogenase
Glyma18g04080	-2285.776	9535.754	26S proteasome regulatory complex, ATPase RPT4
Glyma19g03390	-2344.831	3190.5	Unknown
Glyma19g32940	-2176.029	2579.558	Fatty acid desaturase
Glyma20g24930	-2803.585	6535.602	3-ketoacyl-CoA synthase
Glyma20g32930	-2867.321	2078.549	Cytochrome P450 77A3