The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus

Giang T. H. Vu[a,†,*], Thomas Schmutzer[a], Fabian Bull[a], Hieu X. Cao[a], Jörg Fuchs[a], Trung D. Tran[a], Gabriele Jovtchev[a, ‡], Klaus Pistrick[a], Nils Stein[a], Ales Pecinka[b], Pavel Neumann[c], Petr Novak[c], Jiri Macas[c], Paul H. Dear[d], Frank R. Blattner[a,e], Uwe Scholz[a], Ingo Schubert[a,f,*]

[a]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany; [b]Max Planck Institute for Plant Breeding Research (MPIPZ), Carl-von-Linné-Weg 10, 50829 Köln, Germany; [c]Biology Centre of the Academy of Sciences of the Czech Republic, Institute of Plant Molecular Biology, 370 05, Česke Budejovicé, Czech Republic; [d]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK; [e]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany; [f]Faculty of Science and Central European Institute of Technology, Masaryk University, 61137 Brno, Czech Republic; [†]previous address[b], [‡]present address: Institute for Biodiversity and Ecosystem Research 2 Yurii Gagarin Street, Sofia 1113, Bulgaria.

Received _____.

[*]Corresponding authors (vu@ipk-gatersleben.de or schubert@ipk-gatersleben.de)

Abbreviations: AR, allele ratio; DSB, Double-strand breaks; FISH, fluorescent *in situ* hybridization; Gbp, gigabase pair; GO, gene ontology; HC, high confidence; kbp, kilobase pair; LC, low confidence; LTR, long terminal repeat; MA, Model Averaging; Mbp, megabase pair; MITEs, miniature inverted–repeat transposable elements; MYA, million year ago; SINEs, short interspersed elements; SNP, single nucleotide polymorphism; TE, transposable elements; WGD, whole genome duplication; WGS, whole-genome shotgun;

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Abstract

The C-value paradox remains incompletely resolved after over 40 years, and is exemplified by 2,350-fold variation in genome sizes of flowering plants. The carnivorous Lentibulariaceae genus *Genlisea*, displaying a 25-fold range of genome sizes, is a promising subject to study mechanisms and consequences of evolutionary genome size variation. Applying genomic, phylogenetic and cytogenetic approaches, we uncovered bidirectional genome size evolution within the genus *Genlisea*. The *Genlisea nigrocaulis* genome (86 Mbp) has probably shrunk by retroelement silencing and deletion-biased double-strand break repair, from an ancestral size of 400-800 Mbp to become one of the smallest among flowering plants. The *G. hispidula* genome has expanded by whole-genome duplication (WGD) and retrotransposition to 1,550 Mbp. *G. hispidula* became allotetraploid after the split from the *G. nigrocaulis* clade ~29 MYA. *G. pygmaea* (179 Mbp), a close relative of *G. nigrocaulis,* proved to be a recent (auto)tetraploid. Our analyses suggest a common ancestor of the genus *Genlise*a with an intermediate 1C value (400-800 Mbp) and subsequent rapid genome size evolution in opposite directions. Many abundant repeats of the larger genome are absent in the smaller, casting doubt on their functionality for the organism, whilst recurrent WGD seems to safeguard against the loss of essential elements in the face of genome shrinkage. We cannot identify any consistent differences in habitat or life strategy which correlate with genome size changes, raising the possibility that these changes may be selectively neutral.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Genome sizes bear little relation to the apparent complexity of the organism, in what C.A. Thomas (1971) termed the “C-value paradox” and later was called “C-value enigma” by Gregory (2001). Although much genome size variation is now accounted for by non-coding elements and by duplicated or repetitive sequences, questions remain over the selective advantages of larger or smaller genomes and the mechanisms by which genome sizes change over time.

Genomes expand mainly *via* polyploidization (Soltis and Soltis 1999, Soltis *et al.* 2009, Paterson *et al.* 2012) and repeat amplification (Hawkins *et al.* 2006, Piegu *et al.* 2006, Fedoroff 2012). Polyploidisation may happen as autopolyploidy *via* somatic chromosome doubling, if a mitosis is skipped between two replication cycles, or *via* formation of a restitution nucleus during meiotic divisions (Ramsey and Schemske 1998, De Storme and Geelen 2013). Alternatively, allopolyploidy may occur, if diploid gametes of different species fuse, or a chromosome doubling happens after interspecific hybridisation by fusion of haploid gametes. Duplication of parental chromosome sets is required to ensure in the hybrid organism pairing of homologous chromosomes during meiosis. By sequence loss and other spontaneous mutations, e.g. chromosome rearrangements, the polyploids ‘diploidize’ in the course of evolution (Hegarty and Hiscock 2008, Renny-Byfield *et al.* 2013). Amplification and proliferation of the repetitive elements, based e.g. on insertion of retroelements or on unequal recombination of tandem repeats, contribute significantly to genome expansion (reviewed in (Tenaillon *et al.* 2010, Bennetzen and Wang 2014)). A massive accumulation of long terminal repeat (LTR) retrotransposons during the last few million years accounts for an at least twofold genome size increase in maize (SanMiguel *et al.* 1998) and in a wild relative of rice *Oryza australiensis* (Piegu, et al. 2006). The proliferation of lineage-specific transposable elements (TEs) contributed mainly to genome size variation among closely related *Gossypium* species (Hawkins, et al. 2006).

While it is obvious that polyploidization and accumulation of repeat sequences are main players in genomic expansion, less is known about the mechanism(s) of genome shrinkage. The simplest model for genome shrinkage suggests that the DNA loss via deletion is more frequent than sequence amplification (Hughes and Hughes 1995, Petrov *et al.* 1996, Devos *et al.* 2002, Hu *et al.* 2011, Wang *et al.* 2011, Nam and Ellegren 2012). A high rate of spontaneous loss of

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

nonessential DNA was considered as causing the relatively low proportion of non-genic sequences in small genomes. Indeed, there is evidence for a strikingly higher rate of DNA loss in organisms with small genomes such as *Drosophila* and *Arabidopsis thaliana* compared to those with larger genome size such as mammals or *Nicotiana tabacum* (Kirik *et al.* 2000, Petrov 2001).

Smaller genomes usually have accumulated LTR retrotransposons in their pericentromeric heterochromatin, and few such elements occur within genic regions, while larger genomes reveal a higher number and a more even distribution of retroelements along the chromosomes. A comparison of the ~400-Mbp rice genome and the ~2,400-Mbp maize genome regarding the presence of LTR retrotransposon families revealed approximately the same number of retroelement families in both species, but at least one order of magnitude higher copy numbers per family in the maize genome than in the rice genome (Baucom *et al.* 2009a, Baucom *et al.* 2009b).

For DNA loss, several mechanisms have been suggested such as transposon-mediated excision, replication slippage, and 'illegitimate recombination' (Petrov, et al. 1996, Devos, et al. 2002, Hu, et al. 2011). In our opinion, deletion-biased DSB repair seems to be the most likely cause for genome shrinkage (Kirik, et al. 2000, Puchta 2005), for several reasons: i) DSB repair is an ubiquitous requirement; ii) hypomorphic or hypermorphic mutants of single DSB repair components may result in a bias between repair pathway variants; iii) even a small bias towards either deletions or insertions can have an evolutionary impact; the more so because misrepair events in plant shoot meristems, if viable, may be transferred via germ cells to the next generation; iv) erroneous DSB repair encompasses phenomena such as transposon-mediated excision, replication slippage and illegitimate recombination, the latter for instance via the 'single strand annealing' pathway; and v) chromosome rearrangements are the result of DSB mis-repair and even large interstitial deletions, or translocations resulting in dysploid chromosome number reduction, can be survived if no essential genes are lost (Schubert and Lysak 2011).

In *Arabidopsis thaliana,* previously considered to possess the smallest seed plant genome (157 Mbp, (Bennett *et al.* 2003)), genome evolution has been studied by comparison with the less than two-fold larger genome of *A. lyrata.* This revealed the loss of three (peri)centromeres

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

mediated by dysploid chromosome number reduction (Lysak *et al.* 2006b), a lower content of mobile elements, and an excess of deletions over insertions (Hu, et al. 2011) in *A. thaliana*.

To effectively study genome evolution, comparative analysis of closely related small and larger genomes for whom phylogenetic relationship are well understood can provide hints as to the putative mechanisms of genome alteration over evolutionary timescales. Therefore we have chosen the carnivorous genus *Genlisea* (Lentibulariaceae) with a 25-fold range of genome size and some genomes much smaller than that of *A. thaliana* (Greilhuber *et al.* 2006), thus offering a unique model to address fundamental questions of genome size evolution. Here we apply whole genome sequencing, phylogenetic and cytogenetic approaches in order to elucidate potential reasons for - and consequences of - the observed genome size differences. We focus on *G. nigrocaulis* with half the genome size of *A. thaliana*, on *G. hispidula* with an 18-fold larger genome and on *G. pygmaea,* a close relative of *G. nigrocaulis,* for which we uncovered a recent WGD.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Results and Discussion

Despite having little non-genic DNA, the *G. nigrocaulis* genome displays distinct heterochromatin

We sequenced genomes and transcriptomes of *G. nigrocaulis* (2n = 40; 86 Mbp) and *G. hispidula* (2n = 40; 1,550 Mbp; Fig. 1). For *G. nigrocaulis,* we obtained 54.15 Gbp raw sequences (629.7x genome coverage) by whole-genome shotgun (WGS) sequencing, and assembled 6,968 scaffolds comprising 64.93 Mbp (Tables S2, 3, 4).

Of the ~65 Mbp of assembled *G. nigrocaulis* genomic sequences, 34.02 Mbp (39.6% of the entire genome) is genic DNA including coding parts of mobile elements (Table 1). Dividing the remaining 52 Mbp of the 86 Mbp genome, that include telomeric, centromeric and rDNA sequences, by the chromosome number of *G. nigrocaulis* (n=20), suggests an average of 2.6 Mbp of unique intergenic, and dispersed and clustered repetitive sequences per chromosome. In *A. thaliana* the centromeric and pericentromeric regions alone range from 4 to 9 Mbp per chromosome (Hosouchi *et al.* 2002).

Besides coding and non-coding unique sequences that constitute euchromatin, even the small *G. nigrocaulis* genome with its low repeat content forms detectable heterochromatin structures (Fig. S1), as concluded from i) intensely DAPI-stained chromocenters, ii) DNA and histone H3K27 methylation patterns after immunostaining of nuclei, as well as iii) fluorescent *in situ* hybridization (FISH) with a highly repetitive sequence. The single highly abundant 161 bp tandem repeat of *G. nigrocaulis* is a candidate for centromeric DNA, based on its presence on each chromosome (Fig. S1).

About 22.8 kbp with similarity to the Arabidopsis plastome (154.5 kbp), including 4 out of 88 chloroplast genes, were found interspersed within 24 *G. nigrocaulis* contigs, and 4.1 kbp with similarity to the Arabidopsis mitochondrial genome (366.9 kbp) including fragments of 15 out of 122 mitochondrial genes (EnsemblPlants, http://plants.ensembl.org/; TAIR10 release 18) in 9 genomic contigs (Suppl. 5.1.). In Arabidopsis similar proportions of plastid- (11 kbp) and mitochondrial-derived sequences (~7 kbp, besides a large insertion into chromosome 2 of cv. *Col*) invaded the nuclear genome (The Arabidopsis Genome Initiative 2000). For *G. hispidula* 39.8 kbp of mitochondrial sequences interspersed within 51 contigs with evidence for 26

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

mitochondrial genes and 50 kbp of plastid sequences within 59 WGS contigs, related to 22 plastid genes are detected in the WGS assembly of the nuclear DNA. Thus, the differences in plastid and mitochondrial sequence invasion contribute insignificantly to either genome size variation or heterochromatin formation in the two species.

While even nuclei of the very small genome of *G. nigrocaulis* revealed conspicuous heterochromatin, *G. hispidula* nuclei showed no distinct clusters of heterochromatin-specific marks (Fig.S1B). Dispersed heterochromatic features were previously reported for several medium-sized genomes with a high content of dispersed retroelements and only a moderate clustering of tandem repeats (Houben *et al.* 2003).

The two Genlisea species differ as to their repetitive elements

Repetitive elements were identified and characterized using similarity-based clustering of unassembled sequence reads and the REPEATEXPLORER pipeline (Novak *et al.* 2010, Novak *et al.* 2013). Based on clustering analysis of randomly selected WGS reads, 15.9% (13.7 Mbp) of the *G. nigrocaulis* genome and 64.1% (993.5 Mbp) of the *G. hispidula* genome represent repeats, each covering at least 0.01% of the respective genome (Table 1, Fig. S7).

Similar results were obtained by *k*-mer analysis of the assembled contigs which are larger than 300 bp in both Genlisea species, using the indexes of 21-mers (SI section 7.2). The *k*-mer analysis revealed a very high proportion of unique sequences (83.4%) and a very low proportion of repetitive sequences (16.6%) within the *G. nigrocaulis* genome. In contrast, the genome reference of *G. hispidula* reflects a much higher level of repetitive DNA (63.1%, Fig. S9). When applying the *k*-mer index of *G. nigrocaulis*, we identified only 1% of the highly abundant *k*-mer sequences of *G. hispidula* to be shared with repetitive elements of the *G. nigrocaulis* index (SI section 7.2, Fig. S10). However, the comparative *k*-mer analysis is only suitable to judge gain or reduction of identical sequences. Nucleotide variation makes sequences untraceable for this stringent analysis. When instead a BLASTN analysis with a reduced sequence identity of 95% was applied to search for abundant *G. hispidula* sequences in the *G. nigrocaulis* genome, 56% of the abundant sequences of *G. hispidula* were detected. Nevertheless, the genome of *G. hispidula* contains many repetitive sequences that are apparently not present in *G. nigrocaulis.*

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

These elements may have either entered (and/or proliferated in) the *G. hispidula* genome, or were removed from the *G. nigrocaulis* genome, after the divergence of both species. Overall, 9.7% of *G. nigrocaulis* and 41.6% of *G. hispidula* genomes were annotated as mobile elements, tandem repeats or rDNA (Table 1).

Remarkably, LTR retrotransposons, usually the most abundant repeat class, occupy only 7.3% of the *G. nigrocaulis* genome and show very low copy numbers. Many mobile elements, which are highly abundant, transcribed and probably still transposing in *G. hispidula,* are of low copy number, and apparently suppressed, or even undetectable in *G. nigrocaulis* (Fig. 2, Figs. S7, 8). Thus, also a decline of large-scale insertion, as assumed for pufferfish species (Neafsey and Palumbi 2003), might contribute to genome size reduction in *G. nigrocaulis*. Only a few mobile elements, such as a *Ty1/copia* retroelement of the *Bianca* lineage and a *Ty3/gypsy* element of the *Athila* lineage are relatively abundant in *G. nigrocaulis* and seem to be at least not completely silenced since transcripts are detectable. We aligned RNA-seq data to mobile elements that were identified by Blast2GO analysis of all gene models of *G. nigrocaulis.* Interestingly, among the 219 retroelement candidates of the HC gene models only five captured 60% of all aligned RNA-seq reads and three of the five were assigned to the *Bianca* element (SI section 7.4. and Table S13). Long-lasting suppression puts transposable elements at risk of becoming truncated and eventually extinct *via* deletion-biased DSB repair. This might have happened to several elements in *G. nigrocaulis*. The abundant and active retroelement *Bianca* apparently spread through the *G. nigrocaulis* genome after branching from the *G. hispidula* lineage where it is not detectable. A loss in *G. hispidula* is less likely because of the generally high transposon abundance and activity in this species. *Bianca* and a few other elements counteract genome shrinkage in *G. nigrocaulis* but obviously cannot compensate for silencing of other elements and for progressive DNA loss.

Furthermore, *G. nigrocaulis* and the *G. hispidula* also differ in their putative centromeric and telomeric repeat sequences. The most abundant 161 bp tandem repeat (2.3% of the *G. nigrocaulis* genome), a candidate for centromeric sequence in (Fig. S1), is not found among the genomic reads of *G. hispidula*. The basic telomere repeat conserved in most plants

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

$(TTTAGGG)_n$ is present in *G. nigrocaulis* but lacking in the *G. hispidula* genome, where it is substituted by intermingled (TTTCAGG)n and (TTCAGG)n motifs (Tran et al. 2015).

Higher plant genomes may contain less than 20,000 genes

For *G. nigrocaulis* the gene prediction program AUGUSTUS initially suggested 24,749 gene models which after stringent filtering yielded 15,550 'high confidence' and 1,563 'low confidence' genes (Table 1 and SI section 5.1.). High confidence (HC) genes harboring start and stop codons are present in genomic as well as in transcriptomic sequences and have homology in at least one other plant genome (*A. thaliana*, *Utricularia gibba*, *Solanum lycopersicum* and the other *Genlisea* species). Low confidence (LC) genes do not fulfill one of these criteria. The number of high and low confidence genes together results in less genes than found for *A. thaliana* (27,416; http://www.arabidopsis.org/; (The Arabidopsis Genome Initiative 2000)). For *Utricularia gibba*, from the sister genus of *Genlisea* with an estimated genome size of 82 Mbp, 28,494 genes were predicted utilizing solely AUGUSTUS (Ibarra-Laclette *et al.* 2013). For *G. aurea* a minimal number of 17,755 complete and partial protein encoding genes were estimated from 43.4 Mbp assembled sequences out of the presumed 63.6 Mbp genome (Leushkin *et al.* 2013). Comprehensive studies revealed also for the neotenous monocot *Spirodela polyrhiza* (158 Mbp, (Wang *et al.* 2014)) less than 20,000 genes. The minimum eudicot gene set, based on sequenced genomes of 17 species, was estimated to comprise 7,165 genes which were inferred to descend from 4,585 genes of a common ancestral genome, and the estimated number of ancestral angiosperm genes amounts to about 11,000 to 14,000 (for review see (Li *et al.* 2014)). Thus, gene numbers below 20,000 may occur in small genomes of extant higher plants.

***G. hispidula* is allotetraploid**

In order to compare the minute genome of *G. nigrocaulis* with a large genome of the same genus, we also sequenced genome and transcriptome of *G. hispidula* (2n = 40; 1,550 Mbp; Fig. 1). The genome assembly for *G. hispidula* is based on 24.65 Gbp WGS sequences, and we predicted 42,120 'high confidence' and 21,361 'low confidence' genes (Table 1, SI section 5.1).

The higher fragmentation of the genome assembly of *G. hispidula* led to an increase of partial gene models. To show this we analyzed the 248 core eukaryotic genes (CEG) from *A. thaliana* (Parra et al. 2007). In total 98.4% of the CEGs are observed in *G. nigrocaulis* and 96.9% in *G. hispidula*. In contrast to *G. nigrocaulis* where 80% of the HC genes are estimated to be complete in *G. hispidula* with 39% complete genes a higher fragmentation is observed. The majority of the remaining CEGs represents partial genes classified as low confidence genes (SI section 5.1). The total gene number is more than double of that of *G. nigrocaulis,* suggesting a possible WGD in *G. hispidula.* Genome-wide *k*-mer statistics (Table S12) was performed using full length coding sequences of homologous 'high confidence' gene pairs to further address this hypothesis.

The cleaned WGS reads finally represent a 350-fold genome coverage for *G. nigrocaulis* and 14-fold coverage for *G. hispidula* and were used to build the respective *k*-mer indexes using Tallymer. The indexes were applied to the respective data sets of orthologous high confidence genes to compute their corresponding *k*-mer frequencies. The *k*-mer frequencies of homologous gene pairs were compared. The complete set of analyzed genes revealed an average gene copy number of 0.95 for *G. nigrocaulis* and 1.7 for *G. hispidula* (SI section 6.1)*,* indicating that most of the single copy genes in *G. nigrocaulis* occur twice in *G. hispidula,* and supporting a WGD after these lineages separated.

Furthermore, for both species an intra-species variant detection was performed for high confidence genes to detect heterozygous positions (SI section 6.2). The total number of SNPs was 19,391 for *G. nigrocaulis* and 140,556 for *G. hispidula*. Of the *G. nigrocaulis* SNPs 6,084 (31.38%) and of *G. hispidula* SNPs 91,420 (65.06%) displayed an allele ratio (AR) of 0.4 to 0.6 (Fig. 3A). Assuming an equal heterozygosity level in both self-compatible species, 65.06% of alleles at heterozygous loci in *G. hispidula* with a read frequency of about 1:1 (AR of 0.4-0.6, (Lu *et al.* 2013)) suggest a WGD as a consequence of an interspecific hybridisation.

Additionally, genotype calling and haplotype phasing of individual plants showed that most (11/15) tested homologous genes, present as single copies in *G. nigrocaulis,* have two or more copies in *G. hispidula* (Fig. S6, SI dataset S3). Thus, we assume that the *G. hispidula* lineage underwent allotetraploidisation after its split from the *G. nigrocaulis* lineage (Fig. 1). The same

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

chromosome number in both species indicates dysploid chromosome number reduction in *G. hispidula* or its diploid ancestors. A similar situation was described for Brassicaceae (Lysak, et al. 2006b, Mandakova *et al.* 2010).

The lack of fossils, and possibly elevated DNA mutation rates in the *G. nigrocaulis* clade (Ibarra-Laclette, et al. 2013), do not allow precise dating of the split of the *G. nigrocaulis* and *G. hispidula* lineages. Based on Ks values of 50 nuclear genes and assuming Brassicaceae-like mutation rates, we date the split between *G. nigrocaulis* and *G. hispidula* lineages to ~29 Mya. This age might be overestimated presuming an elevated mutation rate for *G. nigrocaulis* (Ibarra-Laclette, et al. 2013).

Plotting genome size data on a species phylogeny (Fig. 1) suggests an intermediate ancestral genome size (400-800 Mbp) at the basis of the Lentibulariaceae, similar to that found in the oldest genus *Pinguicula*, and an apparently bidirectional genome size evolution within *Genlisea*.

The *G. nigrocaulis* genome probably shrank via a 'deletion bias' during break repair

To identify potential mechanisms of the severe genome shrinkage within the *G. nigrocaulis* lineage, we measured the length distribution of exons and introns. A correlation between the average intron size and genome size has been observed for many related organisms (Vinogradov 1999, Wendel *et al.* 2002). From the *G. hispidula* genome, 1,186 genes with complete intron sequences and orthologs in *G. nigrocaulis* were selected. These gene pairs revealed highly conserved exons. The mean intron size is 123 bp in *G. nigrocaulis* that is 25% less than the 164 bp in *G. hispidula* (Fig. 3B, Table S8). This indicates that genome size differences are not only due to loss of repetitive elements but that a mechanism is acting that influences also the gene space.

Genomes are constantly exposed to DNA breakage due to endogenous and exogenous mutagenic impacts. DSBs, when unrepaired, are lethal for dividing cells. If DSB repair does not restore the pre-break structure, resection of break ends may cause deletions.

Alternatively, insertions may happen e.g. *via* 'synthesis-dependent strand annealing' (gene conversion) and subsequent non-homologous end-joining (Vu *et al.* 2014). The outcome of various erroneous repair pathways may become fixed when passed to the next generation via

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

germ cells. Even minor modifications of components involved in DSB repair, manifested as hypo- or hypermorphic mutants, may cause an imbalance between deletions and insertions and thus, progressive shrinkage or expansion of genomes (Fig. 4B; (Puchta 2005)). Repetitive elements indicative of insertions were rarely found within the introns of both species (13 in *G. hispidula* and 6 in *G. nigrocaulis,* mostly MITEs or SINEs). Therefore, intron size differences are apparently caused by deletions in *G. nigrocaulis* rather than by insertions in *G. hispidula*. A deletion-bias during somatic DSB repair as a presumed reason for genome shrinkage may also influence other genome components which are under no or only weak selection, and could explain a considerable proportion of the genome size difference between the two species by genome shrinkage in the *G. nigrocaulis* clade. Also, for several other groups of organisms with small genomes, a preference of large deletions over insertions was observed, that apparently does not depend on selection for genome size (for review see (Petrov 2001)).

The lower number of genes for DNA metabolic processes in *G. nigrocaulis,* as the most deviating ontology distribution compared to that of *G. hispidula* (Fig. 5, Tables S9, 10), could be one possible reason for the deletion bias in *G. nigrocaulis*. However, this category includes also terms as transposon integration and DNA-templated viral transcription, and thus could also reflect difference in the content of mobile elements.

Whole genome duplication counteracts and facilitates genome shrinkage

Another phenomenon besides retrotransposition that could work against genome shrinkage is WGD. Compared to its close relative *G. nigrocaulis, G. pygmaea* has twice the genome size and chromosome number (179 Mbp, 2n = 80) of *G. nigrocaulis*, and both species share the most abundant tandem repeat and the *Ty1/copia* retroelement *Bianca.* Furthermore, FISH with six unique probes, which labeled two chromosomes in *G. nigrocaulis,* revealed hybridization signals on four chromosomes of *G. pygmaea* each. Thus, our results suggest that a recent WGD occurred in *G. pygmaea* after separation from the *G. nigrocaulis* lineage (Figs. 1, 4A) and mediates tolerance against further even large-scale deletions. Recurrent WGDs (Fig. 4B), might have occurred also in populations of *G. aurea* and *G. repens* (Fig. 1) and possibly in other Lentibulariaceae species with very small genomes. The risk of lethality among the progeny of

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

individuals with very small genomes under progressive deletion-mediated genome shrinkage can be reduced when WGD creates back-up copies of essential genome components as in *G. pygmaea.*

Genome size evolution is divergent within the genus Genlisea

Our comparative studies of one of the smallest known plant genomes (*G. nigrocaulis*) and its congeneric relative with an 18-fold larger genome (*G. hispidula*) reveal an unprecedented case of divergent genome evolution: short introns and a low repeat content indicate genome shrinkage *via* suppression of mobile elements and a presumed preference for deletions during DSB repair in *G. nigrocaulis*, while WGD and possibly retrotransposition led to genome expansion in *G. hispidula*. The relative contribution of these factors to the genome size difference is difficult to estimate, because the actual genome size of the ancestors of *G. hispidula* is unknown. If for both ancestors of *G. hispidula* similar genome sizes (400-800 Mbp) are assumed, WGD-independent genome expansion might have occurred (in case the ancestral genome size was closer to 400 Mbp) or not (in case it was closer to 800 Mbp). In any case however, shrinkage did not occur or was (over)compensated by genome expansion. Assuming a genome size of ~800 Mbp in the last common ancestor and a split between both lineages ~29 Mya (Fig. 1), a DNA loss of on an average 50 bp/1C genome/generation in *G. nigrocaulis* is estimated (SI section 8). If in small genomes DSB repair generates more or larger deletions than in large genomes (Kirik, et al. 2000, Vu, et al. 2014), deletion transmission to the next generation could rapidly reduce genome size, provided the deletions do not include essential genes, or a backup was created before, by recent WGD(s).

Genome size evolution might be neutral in a wide range and many repeats are dispensable

Both investigated species are perennials of similar morphology and live under similar ecological conditions (in nutritient-poor, acidic, waterlogged swampy seepages of tropic regions). Similarly, genome size variation showed no correlation to life strategy diversification within the genus *Allium* (Ohri and Pistrick 2001, Gurushidze *et al.* 2012). Among polyploid *Nicotiana* species

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Leitch et al (2008) observed in five species an increase and in four species a decrease genome size independent of the age of the WGD, and even descendants of the same donor genomes responded differently. Thus, even after WGD no obvious selection pressure regarding the trend of further genome size alteration was detected. The observed large-scale genome shrinkage as well as genome expansion within the genus *Genlisea,* might have been for millions of years a selection-neutral feature (Figs. 1, 4). This assumption gains support from phylogenetic data. The phylogenetic groups of Genlisea species with either small or large-genomes harbor annual and perennial species. Geographic distribution and/or habitat preferences vary similarly in both groups (Fig. 1). Also Fleischmann et al. (2014) could not find any consistent correlation in favor of a plausible selection for genome size alteration within the genus *Genlisea*. Because we cannot see for the respective Genlisea species any obvious factors favoring larger or smaller genomes, we assume the null hypothesis, i.e., that there is no adaptive value for genome size variability within the investigated frame. However, we cannot exclude the possibility that hitherto unidentified factors might select for larger or smaller genomes in these species. Experiments with large populations varying in genome sizes under stable and well defined environmental conditions (habitats), or alternatively with identical genome size within habitats varying as to specifically defined factors might theoretically solve the question of the adaptive value of genome size for specific taxa. However, such experiments seem unfeasible within a reasonable timeframe and effort.

Moreover, the enormous genome shrinkage within *G. nigrocaulis* and other species of this clade, as well as in *U. gibba* (Ibarra-Laclette, et al. 2013), indicates that large proportions of non-coding regions are indeed dispensable. This is not only true for non-coding repetitive sequences of the large genome which are absent from the small genome, but also for the retroelement *Bianca* of *G. nigrocaulis* which is not detectable in the large genome of *G. hispidula.* These observations challenge the current paradigm that most DNA sequences are of functional importance for the carrier organism (see also (Palazzo and Gregory 2014)). Based on our data, we suggest that stochastic WGD may increase, and biased DSB repair may decrease genome size during evolution. Biased DSB repair in either direction could be due to random mutations within one or more components involved in DSB repair. A deletion bias might be

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

caused by more and/or larger deletions (compared to the insertions) in a shrinking genome. Large interstitial chromosome deletions may occur simultaneously together with duplications in the sister chromatid (duplication-deletions), but pure large interstitial deletions occur at about the same frequency (~10% of inducible chromosome rearrangements; (Schubert *et al.* 1994)), yielding a rapid net increase in deletions, Large deletions, under non-selective conditions, can easier explain genome shrinkage than a bias towards single base pair deletions. Together these assumptions offer a reasonable explanation for the C-value paradox/enigma within the investigated genus Genlisea.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Conclusions

We addressed the “C-paradox” by comparing genomes of congeneric species with an unprecedented 18-fold genome size difference, including one of the smallest seed plant genomes. Our analyses suggest a common ancestor of intermediate genome size and genome size evolution in opposite directions with whole genome duplication (WGD) and retrotransposition in one, and retroelement loss and deletion-biased double-strand break repair in the other clade. Genome shrinkage and expansion apparently took place under similar environmental conditions, independent of geographic distribution and life strategy. Therefore we speculate that i) wide variation in genome size might be selectively neutral, ii) many repeats of the larger (but also of the smaller) genome seem to be dispensable in other *Genlisea* genomes, challenging their functionality for the organism, and iii) recurrent whole-genome duplication helps to preserve essential genome elements in the face of long-term genome shrinkage.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Experimental Procedures

Plant material

Plant species used in this study were obtained from the following commercial sources: Carnivorous Plants (http://www.bestcarnivorousplants.com/, Ostrava, Czech Republic): *G. africana, G. aurea, G. hispidula, G. margaretae, G. nigrocaulis, G. pygmaea*; Carnivors and more (http://www.carnivorsandmore.de/., Merzig, Germany): *G. nigrocaulis*, *G. subglabra, G. uncinata*; Die Welt der Fleischfressenden Pflanzen (http://www.falle.de/., Gartenbau Thomas Carow, Nüdlingen, Germany): *G. glandulosissima, G. margaretae* and *G. nigrocaulis* and. Herbarium vouchers of *G. hispidula* (Number: GAT 7858, GAT 7859), *G. nigrocaulis* (Number: GAT 7444, GAT 7445) and G. pygmaea (Number: GAT 23586) were deposited at the IPK Gatersleben.

Genome size determination and cytogenetic experiments

For flow cytometric genome size estimations, leaf tissue of *Genlisea* was chopped together with leaf material of either *Arabidopsis thaliana* 'Columbia' (2C = 0.32 pg, (Bennett, et al. 2003)) or *Raphanus sativus* 'Voran'; IPK gene bank accession number RA 34 (2C = 1.11 pg; (Schmidt-Lebuhn *et al.* 2010)) as internal reference standards in nuclei isolation buffer (Galbraith *et al.* 1983) supplemented with 1 % PVP-25, 0.1 % Triton X-100, DNase-free RNase (50 µg/ml) and propidium iodide (50 µg/ml) according to (Dolezel *et al.* 2007). Measurements were performed using a FACStar^{PLUS} flow sorter (BD Biosciences, New Jersey, USA) and calculations of the genome size were done as described previously (Dolezel, et al. 2007).

Chromosomes were prepared from ethanol:glacial acetic acid (3:1)-fixed young flower buds (*G. nigrocaulis; G. pygmaea*) or young leaves (*G. hispidula*) (for details see SI section 2). For chromosome counting, preparations were stained with 1 µg/ml DAPI in antifade solution (Vectashield, Vector Laboratories).

Immunostaining experiments were performed on flow-sorted 2C leaf nuclei as previously described (Lysak *et al.* 2006a) using the following primary antibodies; mouse anti-5-methylcytosine (Eurogentec), rabbit anti-H3K4me2 and anti-H3K27me1 (Millipore). As secondary antibodies anti-mouse-Alexa 488 and anti-rabbit rhodamine were used, respectively.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Fluorescent in situ hybridizations (FISH) was done according to (Lysak, et al. 2006a). Probes for single copy sequences of *G. nigrocaulis* and for the 161 bp repeat (for primers see Table S1) were prepared by PCR and labeled by nick translation (Lysak, et al. 2006a).

Genome sequencing and assembly

Genlisea species used for this study were identified and their genome size was measured before DNA and RNA was isolated for sequencing (SI sections 1–2). Genomic sequences from libraries with different insert sizes (200 bp – 20 kbp), based on isolated nuclei, were generated on Illumina HiSeq2000 and MiSeq, Roche 454 Titanium (SI section 3.1). Sequence reads used for *de novo* assembly were assembled and scaffolded with CLC Assembly Cell 4.2 (CLC bio, Cambridge, MD) and SSPACE (Boetzer *et al.* 2011). For details see SI sections 3.1–3.5.

Annotation

Gene models were derived from a *Genlisea*-specific training of the AUGUSTUS (Stanke and Morgenstern 2005) pipeline using the RNA-Seq assembled transcriptome of *G. hispidula* and *G. nigrocaulis*. Then the trained AUGUSTUS instance was applied to the respective genome reference to predict gene models with *Genlisea*- specific parameter settings. Furthermore, an OrthoMCL (Li *et al.* 2003) analysis of all AUGUSTUS predictions (proteins of *G. nigrocaulis* and *G. hispidula*) was performed against the protein sequences of *A. thaliana* and *U. gibba* (Ibarra-Laclette, et al. 2013) to look for orthologous groups that support a prediction. OrthoMCL was run as recommended in default settings using the blastp tool for the all to all comparison of protein sequences. Pairwise sequence similarities between protein sequences were calculated using BLASTP with an e-value cut-off $1E^{-5}$. Markov clustering was applied using an inflation value (OrthoMCL parameter -I) of 1.5. The set of high confidence genes comprised gene predictions that have RNA-Seq support and an orthology link either to *A.thaliana*, *U.gibba* or one of the *Genlisea* species. If only one criterion was fulfilled, we validated the quality of the prediction by searching for significant blastp hits (e-value cut-off $1E^{-5}$ and percentage of identity >40%) against a collection of protein sequences of reference plant species (*A. thaliana*, *A. lyrata*, *U. gibba*, *S. lycopersicum* and *Vitis vinifera*) downloaded from the Ensembl Plants (Kersey *et al.* 2012). The Tophat and Cufflinks (Trapnell *et al.* 2010) pipelines were run as additional approaches to confirm the quality of gene predictions and to get better consensus gene

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

predictions. A bi-directional BLAST (blastn) using the coding sequences of *G. nigroculis* and *G. hispidula* was performed to select a suitable set of homologous genes for the intron and exon structure comparison. A set of 1,186 homologous gene pairs of both *Genlisea* species was selected (e-value < $1E^{-30}$, the alignment between the two sequences covered at least 80% of the longer sequence). These genes also showed homology with *U. gibba* sequences (BLASTP, e-value < $1E^{-20}$).

'High confidence' and 'low confidence' genes were annotated using BLAST2GO (Conesa and Gotz 2008). The complete details are described in the SI sections 4–5.

Polyploidy detection

The *k*-mer frequencies of the coding sequences of 1,186 homologous gene pairs of *G. nigrocaulis* and *G. hispidula* were compared. The average copy number is 0.95 for *G. nigrocaulis* and is 1.7 for *G. hispidula.* These values are close to a 1:2 ratio, as expected in case of a WGD event (SI section 6.1). The polyploidy was then confirmed by genome-wide SNP calling within intra-specific transcripts. RNA-Seq reads were aligned with ‘sensitive’ parameter settings by bowtie2 (Langmead and Salzberg 2012) to the set of ‘high confidence’ genes in *G. nigrocaulis* and *G. hispidula, respectively*. The resulting alignments are affiliated into the variant calling process using VCFtools (Danecek *et al.* 2011). Potentially false positive variants were eliminated by discarding variants of inadequate read coverage (<10-fold) and insufficient variant quality (<150). 31.38% of the *G. nigrocaulis* SNPs and 65.06% of *G. hispidula* SNPs were detected with an allele ratio (AR) of 0.4 to 0.6 (SI section 6.2). Assuming an equal heterozygosity level in both self-compatible species, 65.06% of alleles at heterozygous loci with a read frequency of about 1:1 (AR of 0.4-0.6, (Lu, et al. 2013)) support the assumption of allotetraploidy in *G. hispidula*. Furthermore, copy numbers of fifteen randomly selected homologous gene pairs were determined by genotype calling and haplotype phasing. The amplicons of these investigated genes were amplified from three individuals of each species and then Sanger sequenced for identifying interhomeolog or intergenomic polymorphism. Amplicons of *G. hispidula* genes with multiple interhomeolog variants were cloned and Sanger sequenced for sorting variants into haplotypes (SI section 6.3).

Repeat analysis

Repetitive elements were identified using similarity-based clustering of unassembled sequence reads (Novak, et al. 2010) and further characterized using the REPEATEXPLORER pipeline (Novak, et al. 2013) (SI section 7.1). In addition, the SINE-Finder tool (Wenke *et al.* 2011) was used for detection of short interspersed nuclear elements (SINE) in *G. nigrocaulis* and *G. hispidula* genomes. The *K*-mer analysis was done for genome reference sequences of both *G. nigrocaulis* and *G. hispidula* species. 21-mer index libraries were generated from WGS reads of each species and used to analyze the *k*-mer frequencies of the genome reference sequences using Tallymer (Kurtz *et al.* 2008). The *k*-mer frequencies were normalized by the respective sequencing depth (350-fold for *G. nigrocaulis* and 14-fold for *G. hispidula*). Furthermore, to detect the *k*-mers that became reduced or amplified in either species, the shared *k*-mers occurring with high frequency in the genomes of *G. nigrocaulis* and *G. hispidula* were investigated. Strikingly, only 1% of these *k*-mers are shared between both *Genlisea* species, indicating that the genome of *G. hispidula* contains many repetitive sequences that are not present in *G. nigrocaulis.* Additional details and specifications are presented in the SI sections 7.1-7.3.

Ks-based dating

Fifty homologous gene pairs from both species with two homoeologous copies in *G. hispidula* were randomly selected from the high-confidence gene sets and used for age calculation for the split between the lineages of *G. hispidula* and *G. nigrocaulis* and between both parental lineages contributing to the allotetraploid *G. hispidula* (SI section 8). Sequences were aligned with MUSCLE (Edgar 2004) and manually corrected. The pairwise Ks values were calculated with KAKS_CALCULATOR (Zhang *et al.* 2006) using Model Averaging (MA). The approximate age estimates were calculated with a neutral mutation rate of 1.5×10^{-9} mutations per site per year using the formula

$$\text{Age} = \text{Ks} / 2 \times 1.5 \times 10^{-9}.$$

Competing financial interests

The authors declare no competing financial interests.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Author contributions

Conceived and coordinated the study: GTHV and IS. Participated in study design: FRB, US, JF, JM, PHD, AP. Performed the experiments: GTHV, JF, HXC, TDT, GJ, FRB, NS. Genome assembly, RNA-Seq analysis, gene annotation, comparative analysis and polyploidy study: TS, FB, US. Further data analysis: FRB, PNe, PNo, JM, KP, GTHV, HXC, JF.

GTHV, IS, TS, FRB, PHD wrote the manuscript with input from US, JM, FB, JF, AP, NS. All authors read and approved the final manuscript.

Acknowledgments

We thank Klaus Mayer, Munich, Andreas Houben, Renate Schmidt and Florian Mette, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) for helpful comments on the manuscript and Heike Ernst (IPK) for plant photos. This work was supported by IPK, Max-Planck-Institut für Züchtungsforschung, Grant Agency of Czech Republic (P501/12/G090) to JM, Deutsche Forschungsgemeinschaft (SCHU 951/16-1) and the European Social Fund (CZ.1.07/2.3.00/20.0189) to IS and a Ministry of Education and Training (Vietnam) PhD student fellowship to TDT.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Supporting Information

Additional file 1: Supplementary Information: including 8 sections as following with Supplementary Figures 1-11 and Supplementary Tables 1-11

SI section 1. Plant material

SI section 2. Cytogenetics and flow-cytometry

SI section 3. Genlisea genome sequencing and assembly (3.1. Whole-genome shotgun sequencing; 3.2. Quality trimming and error correction of WGS reads; 3.3. *De novo* whole-genome shotgun assembly; 3.4. Scaffolding; 3.5. Post-processing using *k*-mer analysis)

SI section 4. Sequencing and pre-processing of RNA-Seq reads of G. nigrocaulis and G. hispidula

SI section 5. Genome annotation, gene families and comparative genome analysis (5.1. Gene prediction (OrthoMCL analysis of the high confidence gene set, expression analysis of high confidence genes, identification of intron - exon structures); 5.2. Functional Annotation)

SI section 6. Polyploidy (6.1. Polyploidy identified by genome-wide *k*-mer statistics; 6.2. Polyploidy identified by genome-wide SNP calling within transcripts; 6.3. Copy number determination by genotype calling and haplotype phasing in a random gene set)

SI section 7. Repetitive sequences (7.1. Clustering analysis; 7.2. K-mer analsis of the *Genlisea* genome references; 7.3. SINEs in *Genlisea*; 7.4 Mobile elements in RNA-seq data)

SI section 8. Ks-based dating and genome size evolution within Genlisea

Additional file 2: Supplementary Data Set S1. BLAST2GO tables of functional annotation of 'high confidence' and 'low confidence' genes (Excel)

Additional file 3: Supplementary Data Set S2. Average *k*-mer frequency of 1,186 homologous genes supporting polyploidy in *G. hispidula* (Excel)

Additional file 4: Supplementary Data Set S3. Genotype calling and haplotype phasing in a set of randomly selected gene pairs of *G. nigrocaulis* and *G. hispidula.* (Excel)

Additional file 5: Supplementary Data Set S4. Alignment of coding sequences of 50 randomly selected genes used for Ks-based age estimations.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

References

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M., Westerman, R.P., Sanmiguel, P.J. and Bennetzen, J.L. (2009a) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet*, **5**, e1000732.

Baucom, R.S., Estill, J.C., Leebens-Mack, J. and Bennetzen, J.L. (2009b) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Res.*, **19**, 243-254.

Bennett, M.D., Leitch, I.J., Price, H.J. and Johnston, J.S. (2003) Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Ann. Bot.*, **91**, 547-557.

Bennetzen, J.L. and Wang, H. (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.*, **65**, 505-530.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W. (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578-579.

Conesa, A. and Gotz, S. (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*, **2008**, 619832.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. and Durbin, R. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156-2158.

De Storme, N. and Geelen, D. (2013) Sexual polyploidization in plants--cytological mechanisms and molecular regulation. *New Phytol.*, **198**, 670-684.

Devos, K.M., Brown, J.K.M. and Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.*, **12**, 1075-1079.

Dolezel, J., Greilhuber, J. and Suda, J. (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.*, **2**, 2233-2244.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792-1797.

Fedoroff, N.V. (2012) Presidential address. Transposable elements, epigenetics, and genome evolution. *Science*, **338**, 758-767.

Fleischmann, A. (2012) *Monograph of the Genus Genlisea* Poole, Dorset, England: Redfern Natural History Productions.

Fleischmann, A., Michael, T.P., Rivadavia, F., Sousa, A., Wang, W., Temsch, E.M., Greilhuber, J., Muller, K.F. and Heubl, G. (2014) Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Ann Bot*, **114**, 1651-1663.

Fleischmann, A., Schaferhoff, B., Heubl, G., Rivadavia, F., Barthlott, W. and Muller, K.F. (2010) Phylogenetics and character evolution in the carnivorous plant genus *Genlisea* A. St.-Hil. (Lentibulariaceae). *Mol. Phylogenet. Evol.*, **56**, 768-783.

Galbraith, D.W., Harkins, K.R., Maddox, J.M., Ayres, N.M., Sharma, D.P. and Firoozabady, E. (1983) Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science*, **220**, 1049-1051.

Gregory, T.R. (2001) Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biol Rev Camb Philos Soc.*, **76**, 65-101.

Greilhuber, J., Borsch, T., Müller, K., Worberg, A., Porembski, S. and Barthlott, W. (2006) Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biol. (Stuttg.)*, **8**, 770-777.

Gurushidze, M., Fuchs, J. and Blattner, F.R. (2012) The evolution of genome size variation in drumstick onions (*Allium* subgenus *Melanocrommyum*). *Syst. Bot.*, **37**, 96-104.

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A. and Wendel, J.F. (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.*, **16**, 1252-1261.

Hegarty, M.J. and Hiscock, S.J. (2008) Genomic clues to the evolutionary success of polyploid plants. *Curr. Biol.*, **18**, R435-444.

Hosouchi, T., Kumekawa, N., Tsuruoka, H. and Kotani, H. (2002) Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. *DNA Res.*, **9**, 117-121.

Houben, A., Demidov, D., Gernand, D., Meister, A., Leach, C.R. and Schubert, I. (2003) Methylation of histone H3 in euchromatin of plant chromosomes depends on basic nuclear DNA content. *The Plant J.*, **33**, 967-973.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J.D., Ossowski, S., Ottilar, R.P., Salamov, A.A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M.E., Bergelson, J., Carrington, J.C., Gaut, B.S., Schmutz, J., Mayer, K.F., Van de Peer, Y., Grigoriev, I.V., Nordborg, M., Weigel, D. and Guo, Y.L. (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.*, **43**, 476-481.

Hughes, A.L. and Hughes, M.K. (1995) Small genomes for better flyers. *Nature*, **377**, 391.

Ibarra-Laclette, E., Lyons, E., Hernandez-Guzman, G., Perez-Torres, C.A., Carretero-Paulet, L., Chang, T.H., Lan, T., Welch, A.J., Juarez, M.J., Simpson, J., Fernandez-Cortes, A., Arteaga-Vazquez, M., Gongora-Castillo, E., Acevedo-Hernandez, G., Schuster, S.C., Himmelbauer, H., Minoche, A.E., Xu, S., Lynch, M., Oropeza-Aburto, A., Cervantes-Perez, S.A., de Jesus Ortega-Estrada, M., Cervantes-Luevano, J.I., Michael, T.P., Mockler, T., Bryant, D., Herrera-Estrella, A., Albert, V.A. and Herrera-Estrella, L. (2013) Architecture and evolution of a minute plant genome. *Nature*, **498**, 94-98.

Kersey, P.J., Staines, D.M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J.C., Hughes, D.S., Keenan, S., Kerhornou, A., Koscielny, G., Langridge, N., McDowall, M.D., Megy, K., Maheswari, U., Nuhn, M., Paulini, M., Pedro, H., Toneva, I., Wilson, D., Yates, A. and Birney, E. (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91-97.

Kirik, A., Salomon, S. and Puchta, H. (2000) Species-specific double-strand break repair and genome evolution in plants. *EMBO J*, **19**, 5562-5566.

Kurtz, S., Narechania, A., Stein, J.C. and Ware, D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, **9**, 517.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357-359.

Leitch, I.J., Hanson, L., Lim, K.Y., Kovarik, A., Chase, M.W., Clarkson, J.J. and Leitch, A.R. (2008) The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). *Ann. Bot.*, **101**, 805-814.

Leushkin, E.V., Sutormin, R.A., Nabieva, E.R., Penin, A.A., Kondrashov, A.S. and Logacheva, M.D. (2013) The miniature genome of a carnivorous plant *Genlisea aurea* contains a low number of genes and short non-coding sequences. *BMC Genomics*, **14**, 476.

Li, J., Tang, H., Bowers, J.E., Ming, R. and Paterson, A.H. (2014) Insights into the Common Ancestor of Eudicots. In *Adv. Bot. Res.* (Andrew, H.P. ed: Academic Press, pp. 137-174.

Li, L., Stoeckert, C.J., Jr. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178-2189.

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S. and Costich, D.E. (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.*, **9**, e1003215.

Lysak, M., Fransz, P. and Schubert, I. (2006a) Cytogenetic analyses of Arabidopsis. *Methods Mol. Biol.*, **323**, 173-186.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K. and Schubert, I. (2006b) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *PNAS*, **103**, 5224-5229.

Mandakova, T., Joly, S., Krzywinski, M., Mummenhoff, K. and Lysak, M.A. (2010) Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *Plant Cell*, **22**, 2277-2290.

Nam, K. and Ellegren, H. (2012) Recombination drives vertebrate genome contraction. *PLoS Genet.*, **8**, e1002680.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Neafsey, D.E. and Palumbi, S.R. (2003) Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. *Genome Res.*, **13**, 821-830.

Novak, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792-793.

Ohri, D. and Pistrick, K. (2001) Phenology and genome size variation in *Allium* L. - a tight correlation? *Plant Biol. (Stuttg.)*, **3**, 654-660.

Palazzo, A.F. and Gregory, T.R. (2014) The case for junk DNA. *PLoS Genet*, **10**, e1004351.

Parra, G,, Bradnam, K. and Korf, I.(2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* **23,** 1061–1067. doi:10.1093/bioinformatics/btm071

Paterson, A.H., Wendel, J.F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K.C., Shu, S., Udall, J., Yoo, M.J., Byers, R., Chen, W., Doron-Faigenboim, A., Duke, M.V., Gong, L., Grimwood, J., Grover, C., Grupp, K., Hu, G., Lee, T.H., Li, J., Lin, L., Liu, T., Marler, B.S., Page, J.T., Roberts, A.W., Romanel, E., Sanders, W.S., Szadkowski, E., Tan, X., Tang, H., Xu, C., Wang, J., Wang, Z., Zhang, D., Zhang, L., Ashrafi, H., Bedon, F., Bowers, J.E., Brubaker, C.L., Chee, P.W., Das, S., Gingle, A.R., Haigler, C.H., Harker, D., Hoffmann, L.V., Hovav, R., Jones, D.C., Lemke, C., Mansoor, S., ur Rahman, M., Rainville, L.N., Rambani, A., Reddy, U.K., Rong, J.K., Saranga, Y., Scheffler, B.E., Scheffler, J.A., Stelly, D.M., Triplett, B.A., Van Deynze, A., Vaslin, M.F., Waghmare, V.N., Walford, S.A., Wright, R.J., Zaki, E.A., Zhang, T., Dennis, E.S., Mayer, K.F., Peterson, D.G., Rokhsar, D.S., Wang, X. and Schmutz, J. (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature*, **492**, 423-427.

Petrov, D.A. (2001) Evolution of genome size: new approaches to an old problem. *Trends Genet.*, **17**, 23-28.

Petrov, D.A., Lozovskaya, E.R. and Hartl, D.L. (1996) High intrinsic rate of DNA loss in Drosophila. *Nature*, **384**, 346-349.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A. and Panaud, O. (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.*, **16**, 1262-1269.

Puchta, H. (2005) The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *J. Exp. Bot.*, **56**, 1-14.

Ramsey, J. and Schemske, D.W. (1998) Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.*, **29**, 467-501.

Renny-Byfield, S., Kovarik, A., Kelly, L.J., Macas, J., Novak, P., Chase, M.W., Nichols, R.A., Pancholi, M.R., Grandbastien, M.-A. and Leitch, A.R. (2013) Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *The Plant J.*, **74**, 829-839.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.*, **20**, 43-45.

Schmidt-Lebuhn, A.N., Fuchs, J., Hertel, D., Hirsch, H., Toivonen, J. and Kessler, M. (2010) An Andean radiation: polyploidy in the tree genus Polylepis (Rosaceae, Sanguisorbeae). *Plant Biol. (Stuttg.)*, **12**, 917-926.

Schubert, I. and Lysak, M.A. (2011) Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.*, **27**, 207-216.

Schubert, I., Rieger, R., Fuchs, J. and Pich, U. (1994) Sequence organization and the mechanism of interstitial deletion clustering in a plant genome (Vicia faba). *Mutat. Res.*, **325**, 1-5.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K. and Soltis, P.S. (2009) Polyploidy and angiosperm diversification. *Am. J. Bot.*, **96**, 336-348.

Soltis, D.E. and Soltis, P.S. (1999) Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.*, **14**, 348-352.

Stanke, M. and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, **33**, W465-W467.

Tenaillon, M.I., Hollister, J.D. and Gaut, B.S. (2010) A triptych of the evolution of plant transposable elements. *Trends Plant Sci.*, **15**, 471-478.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.

Thomas, C.A. (1971) The genetic organization of chromosomes. *Annu. Rev. Genet.*, **5**, 237-256.

Tran, T.D., Cao, H.X., Jovtchev, G., Neumann, P., Novak, P., Fojtova, M., Vu, G.T.H. Macas, J. Fajkus, J., Schubert, I. and Fuchs, J. (2015) Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus *Genlisea*. *Plant J* accepted

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511-515.

Veleba, A., Bures, P., Adamec, L., Smarda, P., Lipnerova, I. and Horova, L. (2014) Genome size and genomic GC content evolution in the miniature genome-sized family Lentibulariaceae. *New Phytol.*, **203**, 22-28.

Vinogradov, A.E. (1999) Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.*, **49**, 376-384.

Vu, G.T., Cao, H.X., Watanabe, K., Hensel, G., Blattner, F.R., Kumlehn, J. and Schubert, I. (2014) Repair of site-specific DNA double-strand breaks in barley occurs via diverse pathways primarily involving the sister chromatid. *Plant Cell*, **26**, 2156-2167.

Wang, W., Haberer, G., Gundlach, H., Glasser, C., Nussbaumer, T., Luo, M.C., Lomsadze, A., Borodovsky, M., Kerstetter, R.A., Shanklin, J., Byrant, D.W., Mockler, T.C., Appenroth, K.J., Grimwood, J., Jenkins, J., Chow, J., Choi, C., Adam, C., Cao, X.H., Fuchs, J., Schubert, I., Rokhsar, D., Schmutz, J., Michael, T.P., Mayer, K.F. and Messing, J. (2014) The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat. Commun.*, **5**, 3311.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Freeling, M., Pires, J.C., Paterson, A.H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A.G., Park, B.S., Weisshaar, B., Liu, B., Li, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G.J., Bonnema, G., Tang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Jin, H., Parkin, I.A., Batley, J., Kim, J.S., Just, J., Li, J., Xu, J., Deng, J., Kim, J.A., Yu, J., Meng, J., Min, J., Poulain, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M.G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P.J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Sato, S., Sun, S., Kwon, S.J., Choi, S.R., Lee, T.H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Z., Li, Z., Xiong, Z. and Zhang, Z. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.*, **43**, 1035-1039.

Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L. and Senchina, D.S. (2002) Intron size and genome size in plants. *Mol. Biol. Evol.*, **19**, 2346-2352.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117-3128.

Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K. and Yu, J. (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics*, **4**, 259-263.

Accession numbers

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

The NGS resource for both *Genlisea* species is accessible at European Nucleotide Archive under project numbers ‘PRJEB1866’ and ‘PRJEB1867’.

Figure legends

Figure 1. Phylogeny, genome size and life style within the genus *Genlisea,* and chromosomes of *G. nigrocaulis* and *G. hispidula.* Scheme of phylogenetic relationships within Genlisea [modified from (Fleischmann *et al.* 2010) and (Fleischmann 2012)] including 1C genome sizes in Mbp ((Greilhuber, et al. 2006, Fleischmann, et al. 2014, Veleba *et al.* 2014), and own measurements) after taxon names. Two different genome sizes found within *G. aurea* and *G. repens* suggest the occurrence of di- and tetraploid cytotypes. Habitat preferences were compiled from (Fleischmann 2012). The distribution of habitat preferences, life form, and geographic areas among clades shows that none of these traits is correlated with large or small genome size. For 50 randomly selected nuclear genes Ks-based dating (using average Brassicaceae mutation rates) of the split of the lineages leading to *G. hispidula* and *G. nigrocaulis* results with 29 My in more than twice the age calculated by Ibarra-Laclette et al. (2013) for the chloroplast *trn*L-F region (13 My). Using homeologous copies within tetraploid *G. hispidula* for the same gene set and applying identical mutation rates, an age of 20 My is estimated for the split between both parental lineages of *G. hispidula*. This is the maximum age for allopolyploidisation, as, according to the phylogenetic tree, it could have happened any time after the split of both parental lineages and before divergence of *G. hispidula* (tetraploid) and *G. subglabra* (presumed to be tetraploid). *G. nigrocaulis* (red box) and *G. hispidula* (blue box) both with 2n=40 chromosomes.

Figure 2. Repeat composition of *G. nigrocaulis* and *G. hispidula*. Repeat abundance (% of genome size) in both species (left), detail subclasses of Long Terminal Repeat retrotransposons (right).

Figure 3. Allele ratio and intron/exon size comparison between *G. nigrocaulis* and *G. hispidula*. (**A**) The allele ratio distribution of SNPs in *G. nigrocaulis* (red) and *G. hispidula* (blue). SNPs fulfill the criteria of quality score >150 and a minor allele frequency of >0.05. In total there are 19,391 SNPs in *G. nigrocaulis* and 140,556 in *G. hispidula,* while SNPs with an allele ratio of 0.4 to 0.6 are 6,084 (31.38%) in *G. nigrocaulis* and 91,420 (65.06%) in *G. hispidula*, supporting the assumption of a WGD *via* allopolyploidy in *G. hispidula*. **(**B**)** Intron and exon size distribution compared between 1,186 homologous genes of *G. hispidula* and *G.*

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

nigrocaulis. Intron length comparison utilized a subset of 814 gene pairs with at least one intron sequence in both species.

Figure 4. WGD is counteracting genome shrinkage in small *Genlisea* genomes. (**A**) *G. pygmaea* (179 Mbp, 2n = 80 chromosomes) reveals FISH signals on four chromosomes for a single copy probe (*G. nigrocaulis* scaffold 17, position 342935 to 353300) that label two *G. nigrocaulis* chromosomes; the same result was obtained with five other unique sequences (Table S1). (**B**) Model of bidirectional genome size evolution as observed within *Genlisea*. While large genomes evolve through WGD and retrotransposition, deletion-biased DSB repair may result in small genomes. Recurrent WGDs of very small genomes could maintain functionality and prevent reaching a threshold below which deletion-biased DSB repair would drive a species to extinction through loss of essential genome components. Such WGD is also likely for some *G. aurea* populations and for *U. gibba* (Ibarra-Laclette, et al. 2013) (Fig. 1). Alternatively, further shrinkage must be stopped through mutations or gain of gene(s), which reverse deletion-biased DSB repair.

Figure 5. Comparative gene ontology annotation of *G. nigrocaulis* (inner circle) and *G. hispidula* (outer circle) for biological processes. The most severe deviation (black triangle) is represented by much less genes involved in DNA metabolic processes in *G. nigrocaulis.*

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Table 1. Global statistics of the Genlisea genome assemblies and annotation

Species	*Genlisea nigrocaulis*	*Genlisea hispidula*
Estimated genome size (Mbp)	86	1550
# contigs / # scaffolds	17,454 / 6,968	95,804 / -
Sum of contig / scaffold length (Mbp)	60.59 / 64.93	203.80 / -
L50 contig (Kbp) / L50 scaffold (Kbp)	17.4 / 173.7	2.3 / -
N50 contig number / N50 scaffold number	938 / 113	27,310 / -
Predicted genes	24,749	69,894
'High confidence' genes[†]	15,550	42,120
'Low confidence' genes[†]	1,563	21,361
Average exon length (bp)[‡]	271	278
Average intron length (bp)[‡]	123	164
Genic sequences (exons and introns) (Mbp)	34.02	54.78
Defined repetitive sequences (%)	9.7	41.6
Total repeats[§] (%)	15.9	64.1

[†] High confidence (HC) genes harboring start and stop codons are present in genomic as well as in transcriptomic sequences and have homology in at least one other plant genome (*A. thaliana*, *U. gibba*, *S. lycopersicum* and the other *Genlisea* species). Low confidence (LC) genes do not fulfill one of these criteria.

[‡] Because of low coverage and incomplete assembly of the *G. hispidula* genome, several introns could be fragmentary and lead to size underestimation; therefore we used for size comparison the values from 1,186 homeologs selected for completeness in both species. The >2-fold higher gene number in *G. hispidula* is an overestimation as a direct consequence of the lower sequencing depth and hence a colloidal genome assembly with fragmented gene models.

[§] Total repeats were calculated as sum of repeat clusters representing at least 0.01% of the genome.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

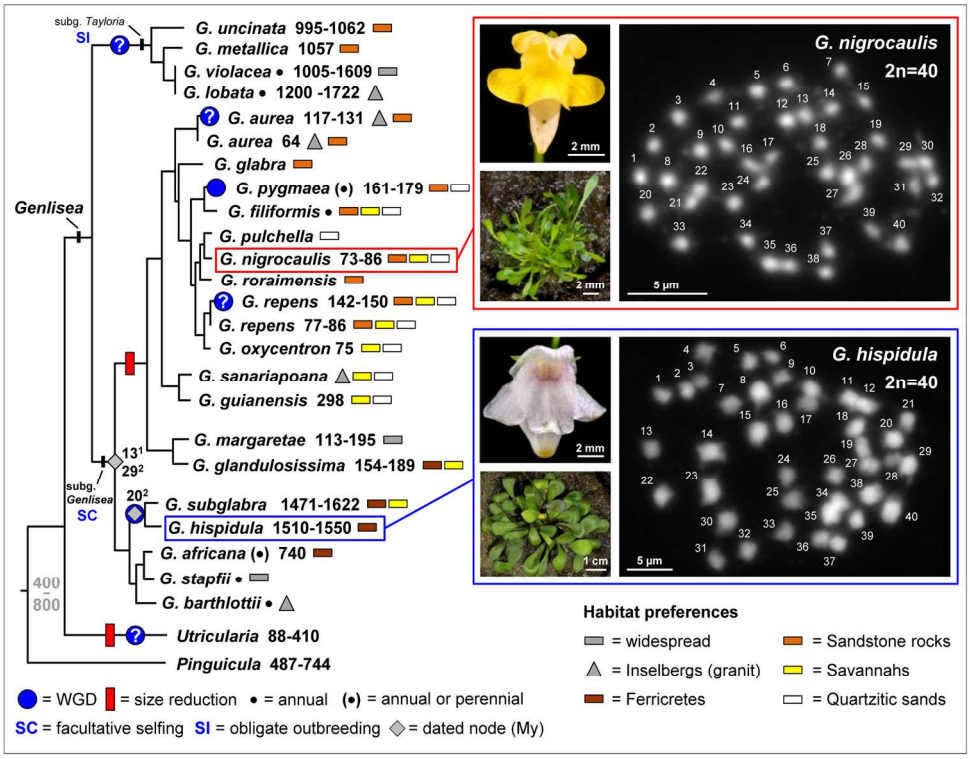


Figure 1. Phylogeny, genome size and life style within the genus Genlisea, and chromosomes of G. nigrocaulis and G. hispidula. Scheme of phylogenetic relationships within Genlisea [modified from (Fleischmann et al. 2010) and (Fleischmann 2012)] including 1C genome sizes in Mbp ((Greilhuber, et al. 2006, Fleischmann, et al. 2014, Veleba et al. 2014), and own measurements) after taxon names. Two different genome sizes found within G. aurea and G. repens suggest the occurrence of di- and tetraploid cytotypes. Habitat preferences were compiled from (Fleischmann 2012). The distribution of habitat preferences, life form, and geographic areas among clades shows that none of these traits is correlated with large or small genome size. For 50 randomly selected nuclear genes Ks-based dating (using average Brassicaceae mutation rates) of the split of the lineages leading to G. hispidula and G. nigrocaulis results with 29 My in more than twice the age calculated by Ibarra-Laclette et al. (2013) for the chloroplast trnL-F region (13 My). Using homeologous copies within tetraploid G. hispidula for the same gene set and applying identical mutation rates, an age of 20 My is estimated for the split between both parental lineages of G. hispidula. This is the maximum age for allopolyploidisation, as, according to the phylogenetic tree, it could have happened any time after the split of both parental lineages and before divergence of G. hispidula (tetraploid) and G. subglabra (presumed to be tetraploid). G. nigrocaulis (red box) and G. hispidula (blue box) both with 2n=40 chromosomes.
170x133mm (300 x 300 DPI)

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

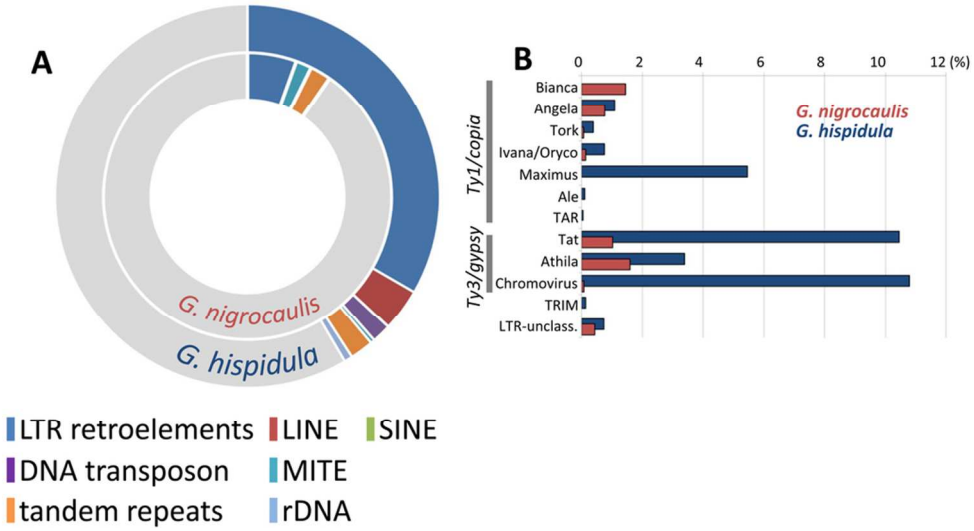


Figure 2. Repeat composition of G. nigrocaulis and G. hispidula. Repeat abundance (% of genome size) in both species (left), detail subclasses of Long Terminal Repeat retrotransposons (right).
89x47mm (300 x 300 DPI)

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

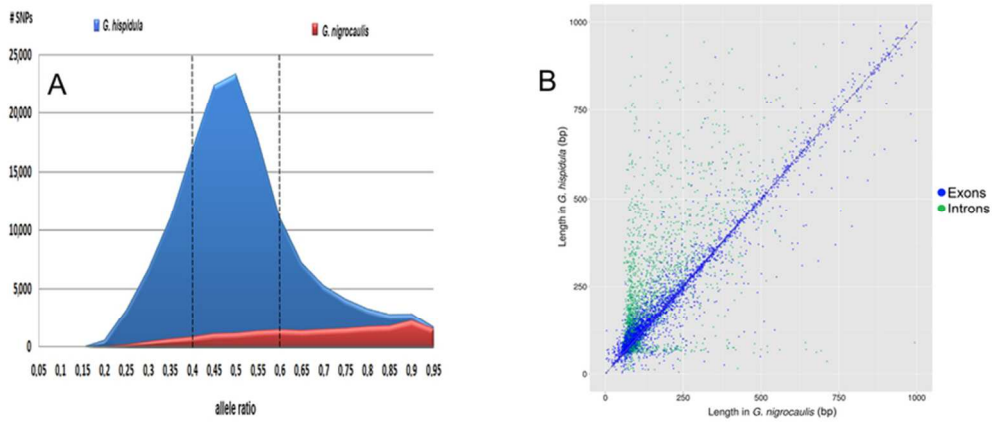


Figure 3. Allele ratio and intron/exon size comparison between G. nigrocaulis and G. hispidula. (A) The allele ratio distribution of SNPs in G. nigrocaulis (red) and G. hispidula (blue). SNPs fulfill the criteria of quality score >150 and a minor allele frequency of >0.05. In total there are 19,391 SNPs in G. nigrocaulis and 140,556 in G. hispidula, while SNPs with an allele ratio of 0.4 to 0.6 are 6,084 (31.38%) in G. nigrocaulis and 91,420 (65.06%) in G. hispidula, supporting the assumption of a WGD via allopolyploidy in G. hispidula. (B) Intron and exon size distribution compared between 1,186 homologous genes of G. hispidula and G. nigrocaulis. Intron length comparison utilized a subset of 814 gene pairs with at least one intron sequence in both species.
73x31mm (300 x 300 DPI)

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

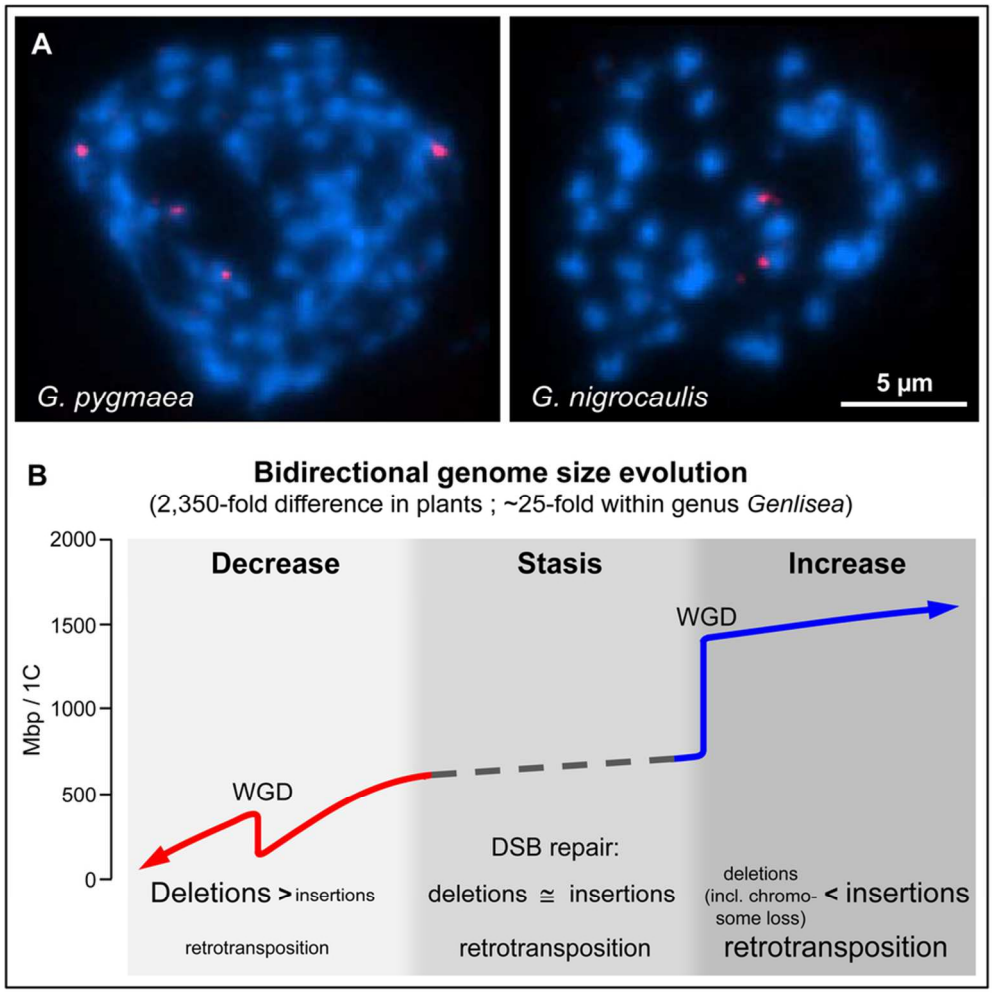


Figure 4. WGD is counteracting genome shrinkage in small Genlisea genomes. (A) G. pygmaea (179 Mbp, 2n = 80 chromosomes) reveals FISH signals on four chromosomes for a single copy probe (G. nigrocaulis scaffold 17, position 342935 to 353300) that label two G. nigrocaulis chromosomes; the same result was obtained with five other unique sequences (Table S1). (B) Model of bidirectional genome size evolution as observed within Genlisea. While large genomes evolve through WGD and retrotransposition, deletion-biased DSB repair may result in small genomes. Recurrent WGDs of very small genomes could maintain functionality and prevent reaching a threshold below which deletion-biased DSB repair would drive a species to extinction through loss of essential genome components. Such WGD is also likely for some G. aurea populations and for U. gibba (Ibarra-Laclette, et al. 2013) (Fig. 1). Alternatively, further shrinkage must be stopped through mutations or gain of gene(s), which reverse deletion-biased DSB repair.
84x84mm (300 x 300 DPI)

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

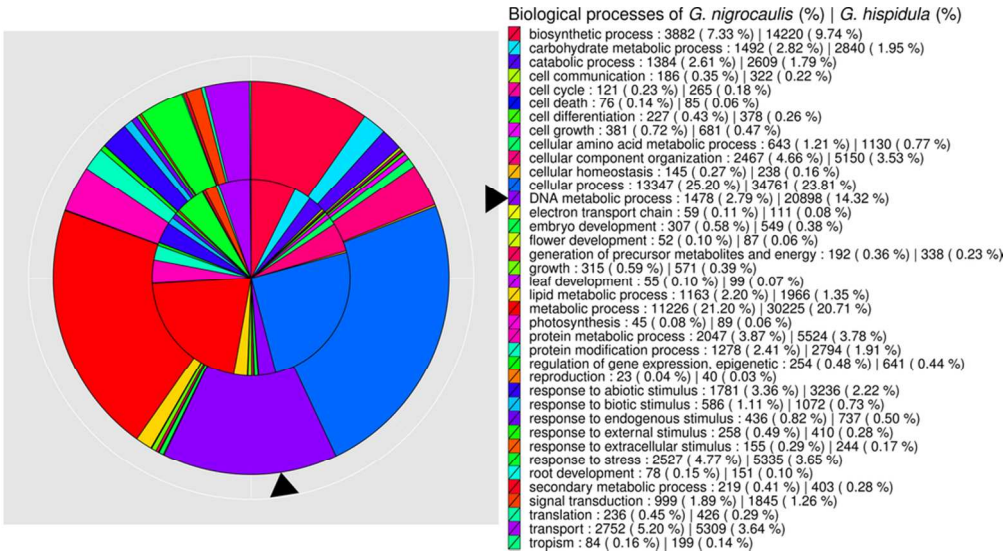


Figure 5. Comparative gene ontology annotation of G. nigrocaulis (inner circle) and G. hispidula (outer circle) for biological processes. The most severe deviation (black triangle) is represented by much less genes involved in DNA metabolic processes in G. nigrocaulis.
92x50mm (300 x 300 DPI)

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Supplementary Materials

Materials and Methods

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

1. Plant material

Plant species used in this study were obtained from the following commercial sources: Carnivorous Plants (http://www.bestcarnivorousplants.com/, Ostrava, Czech Republic): *Genlisea nigrocaulis*, *G. hispidula*, *G. africana*, *G. margaretae*, *G. pygmaea*, *G. aurea*; Carnivors and more (http://www.carnivorsandmore.de/., Merzig, Germany): *G. nigrocaulis*, *G. subglabra, G. uncinata*; Die Welt der Fleischfressenden Pflanzen (Gartenbau Thomas Carow, Nüdlingen, Germany): *G. nigrocaulis*, *G. margaretae, G. lobata* and *G. glandulosissima*. Plants were grown in the greenhouse and determined following the monograph provided by Fleischmann (2012). Vouchers were deposited in the herbarium of the IPK Gatersleben (GAT) (spirit material).

Voucher information

GAT 7444: *Genlisea nigrocaulis* STEYERM.; cultivated in Gatersleben (greenhouse). Origin: Gartenbau Th. Carow, Nüdlingen; Germany; leg. I. Schubert et K. Pistrick 26.08.2011; det. K. Pistrick

GAT 7445: *Genlisea nigrocaulis* STEYERM.; cultivated in Gatersleben (greenhouse). Origin: Carnivors and more Chr. Klein, Merzig, Germany; leg. I. Schubert et K. Pistrick 26.08.2011; det. K. Pistrick

GAT 7858: *Genlisea hispidula* STAPF; cultivated in Gatersleben (greenhouse): 6-8-09. Origin: BestCarnivorousPlants K. Pasek, Ostrava-Poruba, CZ; leg. J. Fuchs et K. Pistrick 09.07.2012; det. K. Pistrick

GAT 7859: *Genlisea hispidula* STAPF; cultivated in Gatersleben (greenhouse): 3; Origin: BestCarnivorousPlants K. Pasek, Ostrava-Poruba, CZ; leg. J. Fuchs et K. Pistrick 09.07.2012; det. K. Pistrick

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

GAT 23586: *Genlisea pygmaea* A.ST.-HIL.; cultivated in Gatersleben (greenhouse): 8/2/11. Origin: BestCarnivorousPlants K. Pasek, Ostrava-Poruba, CZ; leg. et det. K. Pistrick; 30.01.2014

2. Flow-cytometry and cytogenetics

For flow cytometric genome size estimations, roughly 5 mm^2 of leaf tissue of *Genlisea* was chopped with a sharp razor blade together with appropriate amounts of leaf material of either *Arabidopsis thaliana* 'Columbia' (2C = 0.32 pg, (Bennett *et al.* 2003)) or *Raphanus sativus* 'Voran'; IPK gene bank accession number RA 34 (2C = 1.11 pg; (Schmidt-Lebuhn *et al.* 2010)) as internal reference standards in a Petri dish containing 0.6 ml nuclei isolation buffer (Galbraith *et al.* 1983) supplemented with 1 % PVP-25, 0.1 % Triton X-100, DNase-free RNase (50 µg/ml) and propidium iodide (50 µg/ml). The nuclei suspensions were filtered through 35-µm mesh cell strainer caps and stored on ice until measurement. The relative fluorescence intensities of stained nuclei were measured using a FACStar^{PLUS} (BD Biosciences, New Jersey, USA) flow sorter equipped with an argon ion laser INNOVA 90C (Coherent, Palo Alto, CA, USA). Usually, 10,000 nuclei per sample were analyzed. The absolute DNA amounts of samples were calculated based on the values of the G1 peak means. Depending on the availability of material at least two independent measurements per species were performed.

For flow sorting of nuclei, formaldehyde (4%)-fixed leaf material was chopped in LB01 buffer (Dolezel *et al.* 2007) as described above. After staining the nuclei suspension with DAPI (1µg/ml) 2C nuclei were sorted using a FACSAria flow sorter (BD Biosciences, New Jersey, USA) and equivalent amounts of sorted nuclei suspension and sucrose-buffer (Jasencakova *et al.* 2003) were pipetted onto microscopic slides and air-dried overnight.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

For chromosome preparation, young flower buds (*G. nigrocaulis*) or young leaves (*G. hispidula*) were fixed in fixative solution (ethanol:glacial acetic acid = 3:1) for 2 days at room temperature, and stored in 70 % ethanol at 4°C until use. Fixed tissue was washed two times for 5 min in enzyme buffer (10 mM citric acid–sodium citrate, pH 4.5), digested at 37°C for 20 min in enzyme mixture containing 2 % (v/v) cellulase (Duchefa Biochemie), 2 % (w/v) pectolyase (Sigma) and squashed onto microscopic slides in 60 % acetic acid. For chromosome counting the slides were stained with 1 µg/ml DAPI in antifade solution (Vectashield, Vector Laboratories).

Immunostaining and fluorescent in situ hybridization (FISH) were performed as previously described (Lysak *et al.* 2006). For the immunolocalization of methylated DNA a mouse anti-5-methylcytosine antibody (Eurogentec) and for methylated lysine residues of histone H3 rabbit antibodies against H3K4me2 and H3K27me1 (all Millipore) were used. As secondary antibodies anti-mouse-Alexa 488 and anti-rabbit rhodamine were used, respectively.

FISH probes of single copy sequences of *G. nigrocaulis* (Table S1) prepared by PCR and labeled by nick translation were as follows: The FISH probe for the single highly abundant 161 bp repeat was also prepared by PCR (forward: GCCTTATTATGCATCAAATAGCTTC; reverse: GCAATTGGATCCTTTAATAAC-CTC) and labeled by nick translation.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Table S1. FISH probes of single copy sequences of *G. nigrocaulis*

Probe name	Probe location in *G. nigrocaulis* genome				Forward /Reverse primers
	# Scaffold	Start position	End position	Probe size (bp)	
v4c12_p4	72	100,238	108,627	8,389	TGAGTGGTCAAAGAAGACAGGAAG ATTTCCGTTAGCGTAGATTCAAGC
v4c202_p4	17	168,590	177,798	9,208	TTCGATTCTGGATGATAATTGACTG AGTTCAAGCTTCGACGAGTATGTG
v4.2s58_4	58	234,534	243,912	9,378	AGTGATGGAAGTGACTCCAGTGAG TAATTTCGCTCTCTTGCTGCATAC
v4s17_p2	17	342,935	353,300	10,365	ACTCAATCCGGTTCCTGTAAGTTC AGTTCATCCTCTGATGGCCTTAAC
v4s19_p2	19	335,352	343,857	8,505	CCCAGATGAGAGCAATTTGTATTG AACGCATTTCATAGATGAGGATTG
Gps2_p6	2	146,773	157,485	10,712	GCCGAAGCGTCATTTACTCACTAC CAATCCTCTCCAACGCATCTCTTAC

3. *Genlisea* genome sequencing and assembly

3.1. Whole genome shotgun sequencing

30 million and 50 million nuclei from leaf samples of *G. nigrocaulis* and *G. hispidula*, respectively, were flow-sorted to minimize cpDNA contamination. Nuclear DNA was extracted from sorted nuclei using the CTAB method (Clarke 2009). Pure gDNA of *G. nigrocaulis* was subjected to whole genome amplification using REPLI-g Midi kit (Qiagen) in order to get enough DNA for preparing a series of sequencing libraries with different insert size. REPLI-g amplified genomic DNA samples were purified using the QIAamp DNA Mini Kit and quantified by spectrophotometry and pulsed field gel electrophoresis.

Illumina paired-end (PE) and mate-pair (MP) libraries with fragment length ~200 bp (PE), ~2000 bp (PE) and 20 kbp (MP) were prepared from genomic DNA of *G. nigrocaulis*. One additional 454 mate-paired library with 10 kbp fragment length was sequenced. Genomic DNA of *G. hispidula* was sequenced in a single paired-end

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

library with Illumina technology. Sequencing statistics are provided in Table S2. The NGS resource for both *Genlisea* species is accessible at European Nucleotide Archive under project number 'PRJEB1866'.

Table S2. Summary of WGS sequencing raw data

species	library type		ENA accession	reads	sequence (bp)	sequence (Gbp)	est. genome coverage
G. nig	Illumina PE 200 bp (HiSeq)	paired (2x100bp)	ERR296825[a]	2 x 51,313,451	10,262,690,200	10.26	119.3
	Illumina PE 200 bp (HiSeq)	paired (2x100bp)	ERR412882	2 x 3,938,923	787,784,600	0.79	9.2
	Illumina PE 200 bp (HiSeq)	paired (2x100bp)	ERR412881	2 x 191,466,983	38,676,330,566	38.68	449.7
	Illumina MP 4kbp (MiSeq)	paired (2x250bp)	ERR412884	2 x 5,091,981	1,242,185,632	1.24	14.4
	Illumina MP 4kbp (HiSeq)	paired (2x100bp)	ERR412883	2 x 56,408,719	11,281,743,800	11.28	131.2
	Illumina PE 2 kbp	paired (2x100bp)	ERR296826	2 x 6,809,762	1,361,952,400	1.36	15.8
	Roche 454 MP 10 kbp	paired (2x200bp)	ERR299267	2 x 95,006	38,002,400	0.04	0.4
	Illumina MP 20 kbp	paired (2x50bp)	ERR299254	2 x 26,635,295	2,663,529,500	2.66	31.0
∑ *G. nig*					**66,314,219,098**	**66.31**	**771.1**
G. his	Illumina PE 200bp (HiSeq)	paired (2x100bp)	ERR299255[a]	2 x 150,579,430	**30,115,886,000**	**30.12**	**19.4**

[a]Reads of these runs were used for Cluster analysis

3.2. Quality trimming and error correction of WGS reads

To eliminate sequencing errors, we first performed a quality trimming for all constructed WGS data sets and subsequently applied an automatized error correction. All reads are filtered for sequence positions with sufficient quality. For this purpose we applied the quality trimming method of CLC Assembly Cell 4.2 (CLC bio, Cambridge, MD) on paired reads and trimmed ambiguous positions (Phred quality <20). Cutoffs were set on default values requiring a minimal quality of 20 and a good quality fraction for each reads of 0.5. Reads with rejected mates were captured as single reads. Read error correction was performed using Quake (Kelley *et al.* 2010). Therefore an 18-mer index was constructed using Jellyfish (Marcais and Kingsford 2011) which is used by Quake to correct sequencing errors. For both *G. nigrocaulis*

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

and *G. hispidula* the 200 bp (PE) libraries were used for the 18-mer index construction which subsequently was applied to all WGS data sets. After discarding ambiguous positions in the WGS data we derived a 630-fold sequencing coverage of the *G. nigrocaulis* genome and a 15.9-fold coverage of the *G. hispidula* genome. Details are listed in Table S3.

Table S3. Statistics of pre-processed WGS sequence data

species	library type	reads	avg. read length (bp)	sequence (bp)	sequence (Gbp)	est. genome coverage
G. nig	Illumina PE 200 bp (HiSeq)	2 x 37,508,870	91.3	6,774,901,862	6.77	78.8
		1,273,364	81.5	103,995,410	0.10	0.1
	Illumina PE 200 bp (HiSeq)	2 x 3,647,474	98.7	719,712,016	0.72	8.4
		222,363	88.7	19,730,109	0.02	0.2
	Illumina PE 200 bp (HiSeq)	2 x 164,442,874	96.2	33,692,640,296	33.69	391.8
		21,230,841	84.9	466,475,345	0.47	5.4
	Illumina MP 4kbp (MiSeq)	2 x 3,163,801	157.7	998,098,363	1.00	11.6
		797,828	213.0	169,978,068	0.17	2.0
	Illumina MP 4kbp (HiSeq)	2 x 49,062,550	97.6	9,581,211,021	9.58	111.4
		5,493,289	84.9	466,475,345	0.47	5.4
	Illumina PE 2 kbp (HiSeq)	2 x 4,304,184	92.3	794,910,517	0.79	9.2
		915,045	92.2	84,409,780	0.08	1.0
	Roche 454 MP 10 kbp	2 x 90,091	161.8	28,740,568	0.03	0.3
		76,752	206.5	15,845,742	0.02	0.2
	Illumina MP 20 kbp (HiSeq)	2 x 862,077	47.8	82,520,079	0.08	1.0
		3,186,059	47.6	151,708,088	0.15	1.8
∑ ***G. nig***		**559,359,383**	**96.8**	**54,151,352,609**	**54.15**	**629.7**
G. his	Illumina PE 200 bp (HiSeq)	2 x 123,056,614	97.2	23,911,470,917	23.91	15.4
		7,552,333	95.1	736,778,420	0.74	0.5
∑ ***G. his***		**253,665,561**	**97.2**	**24,648,249,337**	**24.65**	**15.9**

3.3. *De novo* whole-genome shotgun assembly

Nuclear DNA PE reads have been assembled with CLC assembly cell (v4.2) to build a contig set with minimal contamination, since purified DNA from extracted nuclei has been used for sequencing. For *G. nigrocaulis* two other libraries of DNA isolated from entire plants were sequenced to gain a complete genome representation. All

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

read libraries were processed in a combined CLC assembly using standard parameter settings and allowing contigs with minimal length of 300 bp.

To establish a quality reference assembly, various control mechanisms were implemented to sort out possible contamination. First we anchored the constructed contigs against contamination free data resources. For that purpose, the assembly of nuclear DNA of *G. nigrocaulis* and of *G. hispidula* and the genome reference assembly of the close related species *Utricularia gibba* served as control. For *G. nigrocaulis* 6,915 contigs were identified as non-contaminated, comprising in total 52 Mbp of cumulative contig size. As an additional control we used RNA-Seq reads from flower tissues. PE reads from flower showed in pre-screenings almost no microbial contamination. RNA-Seq reads were mapped as spliced alignment with Tophat (Kim *et al.* 2013) to the genome reference contigs to capture all contigs with expressed RNA-Seq evidence. Furthermore, all non-anchored WGS contigs were blastn (Altschul *et al.* 1990) analyzed against the NCBI nucleotide collection and checked for hits to angiosperms (e-value cut-off $1E^{-5}$). Further quality in the assembly was gained by removing miss-assembled contigs revealed by coverage analysis. Since erroneous reads might form arbitrary contigs which usually are small and can be detected because of very poor read coverage support, we removed contigs with an average coverage less than 30x from our reference set. Applying the previously explained methods we further integrate 7.28 Mbp (8.234 contigs) to the 52 Mbp of reference sequence. By using the program IMAGE of the PAGIT toolkit (Swain *et al.* 2012), we performed a reference error correction and gap closing. To correct the reference, only reads from nuclear DNA sequence were used as input. IMAGE was run in default parameters for 12 iterations with four gradually decreased *k*-mer sizes (91, 71, 51 and 31) modified after each third iteration step. In total, 1,288

gaps have been closed, decreasing the proportion of 'N' bases in the reference to zero and increasing the cumulative contig size from 59.28 to 60.59 Mbp (17,457 contigs). The constructed WGS assembly reference comprised an L50 of 17,420 bp, and enriched the quality of subsequent scaffolding.

The genome sequence of *G. hispidula* was assembled using the purified DNA from extracted nuclei. We build a WGS assembly using CLC assembly cell (v4.2) with standard parameter settings, allowing contigs with minimal length of 300 bp. The total length of the WGS assembly was 314.5 Mbp, comprised by 303,873 contigs.

3.4. Scaffolding

For scaffolding in *G. nigrocaulis* small and accurate short mate pair libraries (2 kbp) were used first, then mate pair libraries with increasing distances (4 kbp, 10 kbp and 20 kbp) were integrated in a step-wise manner. We applied SSPACE (Boetzer *et al.* 2011) as scaffolding program. Standard parameters were used and mate distances were calculated with CLC assembly cell. The first phase of scaffolding resulted in 12,217 WGS scaffolds with an L50 scaffold size of 117 kbp and a cumulative scaffold size of 64.88 Mbp. Scaffolding integrated many 'N'-gaps that were reduced by a subsequent gap filling approach using GapFiller (Boetzer and Pirovano 2012). In a successive scaffolding run the final reference of 11,468 scaffolds was established, comprising a cumulative scaffold length of 66.5 Mbp for *G. nigrocaulis*.

3.5. Post-processing using *k*-mer analysis

We applied a *k*-mer analysis to the constructed genome reference to detect assembly artifacts which may result from sequencing errors. We used all PE reads of *G. nigrocaulis* that align to the constructed reference as basis for the construction of a *k*-mer index using Tallymer (Kurtz *et al.* 2008). This index was applied to the reference sequence of *G. nigrocaulis* to compute *k*-mer frequencies for each

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

constructed scaffold. The integrated reads represent the genome with a 350-fold coverage. Assembly artifacts that are caused by sequencing errors have a lower *k*-mer frequency which is a consequence of insufficient representation of reads by *k*-mers. Furthermore very small scaffolds also can be an artifact of the assembly process. Therefore we removed short scaffolds (<500bp) with a deficient *k*-mer frequency (*k*-mer frequency <1) and no detectable genes. Final contigs of *G. hispidula* were analyzed with the same approach using the respective data set of WGS sequencing for index construction and excluding artifacts of short contigs (<1000bp) with deficient *k*-mer frequencies. The final WGS assembly of *G. hispidula* with a total length of 203.8 Mbp comprised 95,804 contigs (Table S4).

Table S4. Assembly and scaffolding statistics

	G. nigrocaulis			***G. hispidula***
	contigs all	scaffolds all	scaffolds final	contigs final (>1 kbp)
# contigs/scaffolds	17,454	11,468	6,968	95,804
# contigs/scaffolds > 10kbp	1,580	568	568	185
# contigs/scaffolds > 50kbp	117	315	315	0
cum. contig/scaffold size (Mbp)	60.59	66.71	64.93	203.80
largest contig/scaffold (kbp)	122.6	628.4	628.4	25.7
L50 (kbp)	17.4	162.7	173.7	2.3
N50	938	118	113	27,310

4. Sequencing and pre-processing of RNA-Seq reads of *G. nigrocaulis* and *G. hispidula*

RNA samples from three different organs (traps, leaves and flowers) of *G. nigrocaulis* and *G. hispidula* were isolated using an RNeasy Kit (Qiagen). To annotate transcriptionally active regions in the WGS contigs of *G. nigrocaulis* and *G. hispidula,* deep RNA sequence data (RNA-Seq) were produced by Illumina paired-end sequencing (2x100 bp reads, 200 bp insert size) for all three organs of both

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Genlisea species (Table S5). The RNA-Seq resource is accessible at European Nucleotide Archive under project number 'PRJEB1867'.

Table S5. Statistics of RNA-Seq raw data

species	library type	organ	ENA accession	reads	read length (bp)	sequence (bp)	sequence (Mbp)
G. nig	Illumina PE 200 bp	leaf	ERR299256	2 x 30,072,857	101	6,074,717,114	6,075
	Illumina PE 200 bp	trap	ERR299258	2 x 27,571,185	101	5,569,379,370	5,569
	Illumina PE 200 bp	flower	ERR299260	2 x 32,383,215	101	6,541,409,430	6,541
	Illumina PE 200 bp	plant	ERR299261	2 x 20,791,438	101	4,158,287,600	4,158
∑ *G. nig*						**22,343,793,514**	**22,344**
G. his	Illumina PE 200 bp	leaf	ERR299262	2 x 53,687,723	101	10,844,920,046	10,845
	Illumina PE 200 bp	trap	ERR299264	2 x 59,522,012	101	12,023,446,424	12,023
	Illumina PE 200 bp	flower	ERR299266	2 x 39,037,073	101	7,885,488,746	7,885
∑ *G. his*						**30,753,855,216**	**30,754**

All RNA-Seq data sets were trimmed for sufficient sequence quality (Phred quality >20) using methods of CLC Assembly Cell 4.2 (CLC bio, Cambridge, MD) on paired reads (Table S6). Cutoffs were set on default requiring a minimal quality of 20 and a good quality fraction for each reads of 0.5. Reads with rejected mates are captured as single reads. Read error correction was performed using Quake. For both, *G. nigrocaulis* and *G. hispidula,* the purified sequence datasets of nuclear DNA (WGS PE 200 bp nuclei) were used for the 18-mer index construction with Jellyfish. These constructed indices subsequently were applied for error correction to all RNA-Seq data sets individually using Quake. We observed high contamination rates especially in the data sets from trap-derived sequences. For purification we anchored all RNA-Seq data sets to the quality filtered genome reference of each *Genlisea* species by performing a blastn comparison.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Table S6. Statistics of quality enriched RNA-Seq reads

		quality trimmed					genome mapped			
species	tissue	reads	avg. read length (bp)	sequence (bp)	sequence (Mbp)	%	reads	sequence (bp)	sequence (Mbp)	%
G. nig	leaf	2 x 30,451,762	91.9	5,598,458,263	5,598	92	2 x 24,395,136	4,467,280,890	4,467	74
	trap	2 x 24,292,975	92.4	4,490,884,037	4,491	81	2 x 11,487,196	2,116,838,222	2,117	38
	flower	2 x 24,002,922	84.6	4,059,795,101	4,060	62	2 x 23,414,812	3,958,888,058	3,959	61
	plant	2 x 15,627,508	87.5	2,733,429,174	2,733	66	2 x 14,657,417	2,562,226,624	2,562	62
∑ *G. nig*				**16,882,566,575**	**16,883**			**13,105,233,794**	**13,105**	
G. his	leaf	2 x 31,227,511	93.8	5,856,490,071	5,856	54	2 x 26,316,455	4,929,984,009	4,930	45
	trap	2 x 37,392,797	92.8	6,945,792,547	6,946	58	2 x 31,506,822	5,840,450,660	5,840	49
	flower	2 x 30,367,202	86.3	5,242,371,529	5,242	66	2 x 25,802,728	4,445,903,719	4,446	56
∑ *G. his*				**18,044,654,147**	**18,045**			**15,216,338,388**	**15,216**	

5. Genome annotation, gene families and comparative genome analysis

5.1. Gene prediction

The challenge of an accurate and complete gene prediction was directed towards high accuracy of the prediction approach. This might discard partial or weak prediction and in consequence the total number of gene models might be underestimated. Since we aimed to compare the two *Genlisea* species we focused on the prediction of complete gene models. To establish a gene set in both species, we performed a *Genlisea*-specific training of the AUGUSTUS (Stanke and Morgenstern 2005) pipeline using the RNA-Seq assembled transcriptome of *G. hispidula* and *G. nigrocaulis*. Then the trained AUGUSTUS instance was applied to the respective genome reference to predict gene models (24,749 in *G. nigrocaulis* and 69,894 *G. hispidula*) with *Genlisea*- specific parameter settings.

To select relevant genes and discard possible false positive models we ran several post-processing steps to further enrich the quality of the gene sets. For all gene models with predicted transcription start and end sites we performed a read mapping of our RNA-Seq datasets to the AUGUSTUS predictions. To exclude program biases for the read mapping, we used two independent approaches, the STAR (Dobin *et al.* 2013) pipeline and the Tophat pipeline to select for gene models that have mapping support by our RNA-Seq data. Furthermore, an OrthoMCL (Li *et al.* 2003) analysis of all AUGUSTUS predictions (proteins of *G. nigrocaulis* and *G. hispidula*) was performed against the protein sequences of *A. thaliana* and *U. gibba* [42] to look for orthologous groups that support a prediction. OrthoMCL was run as recommended in default settings using the blastp tool for the all to all comparison of protein sequences. The set of high confidence genes comprised gene predictions that have RNA-Seq support and an orthology link either to *A. thaliana*, *U. gibba* or one of the *Genlisea* species. All other predictions are categorized as candidates for low confidence genes and were further analyzed by various criteria. First we selected predictions with start and stop codon and sequence similarity in the OrthoMCL analysis. These similarities might not fulfill criteria for a link to a known gene of the reference species, but may serve as quality control for a prediction. Gene models that meet both criteria were considered as low confidence genes. If only one criterion is fulfilled, we validated the quality of the prediction by searching for significant blastp hits (e-value cut-off $1E^{-5}$ and percentage of identity >40%) against a collection of protein sequences of reference plant species (*A. thaliana*, *A. lyrata*, *U. gibba*, *S. lycopersicum* and *V. vinifera*) downloaded from EnsemblPlants (Kersey *et al.* 2012).

To confirm the quality of our gene predictions, we ran an independent approach based on RNA-Seq data and the Tophat and Cufflinks (Trapnell *et al.* 2010) pipeline.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

The results obtained by both methods were highly concordant and the number of newly identified genes was very low. For *G. nigrocaulis* only 105 (0.6%) additional gene models (44 high confidence and 61 low confidence genes) were detected that were not overlapping with our gene set. For *G. hispidula* 6,120 (9.6%) additional gene models (3,791 high confidence and 2,329 low confidence genes) have been detected. These additional genes were integrated into our gene sets. To define high confidence we used only the most significant gene models with highest transcript score and FPKM (fragments per kilobase of exon per million fragments mapped) values >10. By protein sequence blast (minimal identity threshold of 40%) we selected high confidence genes corresponding to known proteins from databases of *A.thaliana*, *U.gibba* or *S. lycopersicum*. Cufflinks predictions with no significant protein hit but with high FPKM values >50 were integrated as low confidence genes. The complete gene set of *G. nigrocaulis* comprises 17,113 gene models. Of these, 15,550 (90.87%) are classified as high confidence genes and 1,563 (9.13%) as low confidence genes. The gene set of *G. hispidula* is represented by 63,481 gene models; 42,120 (66.35%) classified as high confidence genes and 21,361 (33.65%) as low confidence genes. We expect that the low sequencing coverage and, as a direct consequence, the incompleteness of the genome assembly in *G. hispidula* led to a higher fragmentation of gene models. Thus, we conclude that the gene number in *G. hispidula* is overestimated. To estimate the completeness of predicted gene models we performed a comparative BLASTX analysis against 248 known core eukaryotic genes (CEGs) from *A. thaliana* (Parra et al. 2007). High stringency was applied with an identity threshold of >40% at amino acid level. Applying a minimal length threshold of 150bp we detected 244 (98.4%) CEGs in *G. nigrocaulis* and 238 (96.0%) *G. hispidula.* To estimate the proportion of complete CEGs in high

confidence genes of both Genlisea species we applied the CEGMA pipeline parameter (at minimum 75% of the CEG length covered). With 199 (~80%) in *G. nigrocaulis* and 97 (~39%) in *G. hispidula* the higher fragmentation in gene models of *G. hispidula* is illustrated. Decreasing the required identity threshold to 10% (used in CEG pipeline) further 33 were found in the HC gene set of *G. nigrocaulis* and 44 in *G. hispidula*. Decreasing the parameter setting to 20% minimal length (>150bp) and screening the low confidence gene sets we found 11 and 97 of the remaining CEGs in *G. nigrocaulis* and *G. hispidula*.

We analyzed *G. nigrocaulis* and *G. hispidula* assemblies for mitochondrial and plastid sequences. The *A. thaliana* mitochondrial and chloroplast genome and the genes assigned to these organellar genomes (see http://plants.ensembl.org/; TAIR10 release 18) were used for comparison. To prevent contamination by *G. nigrocaulis* organellar DNA, we used the WGS sequencing dataset of *G. nigrocaulis* derived from isolated nuclei and performed a WGS assembly of nuclear PE reads (Table S2; ERR296825). A blastn comparison (70% identity and minimal alignment length of HSP of 100 bp) detected 4.1 kbp with similarity to the mitochondrial genome interspersed within 9 WGS contigs and related to 15 (out of 122) mitochondrial genes and 22.8 kbp with similarity to the plastome of *A. thaliana* interspersed within 24 contigs and related to 4 (out of 88) chloroplast genes in the G*. nigrocaulis* genome. In *Arabidopsis* similar proportions of plastid- (11 kbp) and mitochondrial-derived sequences (~7 kbp, besides a large insertion into chromosome 2 of cv. *Col*) invaded the nuclear genome (The Arabidopsis Genome Initiative 2000). We performed a blastn comparison with identical parameter settings for *G. hispidula* and detected 39.8 kbp of mitochondrial sequences interspersed within 51 contigs with evidence for 26 mitochondrial genes and 50 kbp of plastid

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

sequences within 59 WGS contigs, related to 22 plastid genes in the WGS assembly of nuclear DNA.

5.1.1. OrthoMCL analysis of the high confidence gene set

To further support our gene prediction, we clustered proteins of *G. nigrocaulis, G. hispidula, U. gibba* and *A. thaliana* and formed 20,443 orthologous groups (Fig. S2, Table S7). Among these, both *Genlisea* species share 8,749 clusters, including 10,072 (64.8%) *G. nigrocaulis* 'high confidence' genes. Two hundred fifty four clusters containing 822 (5.3%) genes of *G. nigrocaulis,* and 3,320 clusters containing 13,492 genes (32%) of *G. hispidula* were not found in the other three species.

We used OrthoMCL (software version 2.0.3) to define gene family clusters for *G. nigrocaulis, G. hispidula*, *U. gibba, A. thaliana* and *S. lycopersicum* gene models. Protein sequences of *A. thaliana* and *S. lycopersicum* were downloaded from the Ensembl Plants (Kersey, et al. 2012). Pairwise sequence similarities between protein sequences were calculated using BLASTP with an e-value cut-off $1E^{-5}$. Markov clustering was applied using an inflation value (OrthoMCL parameter -I) of 1.5. In total, we identified for *G. nigrocaulis* 12,874 protein sequences with an orthologous sequence in at least one other species (10,259 to *G. hispidula*, 11,424 to *U. gibba*, 10,490 to *A. thaliana*, and 11,019 to *S. lycopersicum*). For *G. hispidula* we detected 12,181 proteins with an orthologous sequence in other species (10,431 to *G. nigrocaulis,* 10,625 to *U. gibba,* 9,595 to *A. thaliana* and 10,137 to *S. lycopersicum*). We presume that the low number of *G. hispidula* sequences found in orthologous pairs is caused by the higher fragmentation of the *G. hispidula* gene set, leading to a high number of singletons. We found that 6,869 orthologous groups contained sequences from all five species and 400 were found to be specific to Lentibulariaceae genomes (*G. nigrocaulis, G. hispidula* and *U. gibba*).

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Table S7. Construction of orthologous groups of protein encoding sequences

Species	Proteins	Proteins in ortholgous groups	Proteins in orthologous groups with at least one other species	Species-specific proteins
A. thaliana	35,286	30,802	23,420	7,382
G. nigrocaulis	15,550	13,623	12,874	749
G. hispidula	42,120	25,487	12,181	13,306
U. gibba	28,032	18,860	16,086	2,774
S. lycopersicum	34,675	24,963	18,545	6,418

5.1.2. Expression analysis of high confidence genes

We studied the different expression of genes in three organs (traps, leaves and flowers) of *G. nigrocaulis* and *G. hispidula* by analyzing the quality enriched RNA-Seq datasets (Table S6 "genome mapped" RNA-Seq reads). We utilized Tophat to align RNA-Seq reads to the high confidence gene set. In addition these read alignments were analyzed by Cufflinks to compute differences in expression levels by analyzing the fragments per kilobase of transcript per million mapped reads (FPKM) in high confidence genes. Genes with an FPKM <5.0 were discarded. In *G. nigrocaulis* we found in total 15,176 expressed genes (97.6%), of which 498 revealed organ-specific expression. In *G. hispidula* 37,480 genes (89.0%) were found to be expressed and 5,440 of these in an organ-specific manner (Fig. S3).

We assume that the latter number is an overestimation because the total amounts of RNA-Seq used in this comparison between *G. nigrocaulis* (10.5 Gbp) and *G. hispidula* (15.2 Gbp) represent different levels of sequencing depth per gene. For the polyploid genome of *G. hispidula* the depth per gene was lower. This may lead to a general underrepresentation of genes in the *G. hispidula* data set and to an increased number of genes captured in one organ only.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

5.1.3. Identification of intron - exon structures

To compare the intron/exon structure of genes of both Genlisea species, we selected a set of 1,186 homologous gene pairs (Fig. 3B, Fig. S4). The lower sequencing depth for the G. hispidula genome led to a higher fragmentation in the corresponding assembly. Therefore, we restricted the comparison of the intron and exon structure to highly similar coding sequences of G. nigrocaulis and G. hispidula. To detect such genes, we used the coding sequences predicted by the AUGUSTUS pipeline. A bi-directional BLAST (blastn) between the two species was done, restricting the hits by an e-value below $1E^{-30}$. Candidate pairs were formed by sequences which mutually met each other as best hit. A candidate pair was considered as homologous if the alignment between the two sequences covered at least 80% of the longer sequence. Of the resulting 1,214 homologous gene pairs, 1,186 were again detected when the published peptides of U. gibba were compared (BLASTP) against the presumed peptide sequences of the 1,214 G. nigrocaulis genes with an e-value below $1E^{-20}$. These 1,186 homologs were further investigated.

Table S8. Exon and intron size statistics for 1,186 homologous gene pairs

	G. nigrocaulis	*G. hispidula*
Max. intron size	1,527	2,124
Med. intron size	83	93
Mean intron size	123.4	163.7
Min. intron size	51	45
Max. exon size	7,316	7,929
Med. exon size	146	151
Mean exon size	270.7	277.9
Min. exon size	41	41

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

The comparison of intron and exon size (Table S8) of the 1,186 homologous genes was based on the gene model structures predicted by the AUGUSTUS pipeline. The analysis of intron size was limited to genes (814) which in both species contained at least one intron. On average *G. nigrocaulis* genes displayed 4.8 and their *G. hispidula* homologues 4.6 introns. We utilized a custom greedy algorithm to assign homologous introns and exons. For each homologous gene pair corresponding introns and exons were assigned based on the alignment algorithm provided by the BioJava framework (v 3.0.8, (Prlic *et al.* 2012)). It uses the standard BLAST substitution matrix (available at ftp://ftp.ncbi.nih.gov/blast/matrices/NUC.4.4) with a gap opening cost of 5 and gap extension cost of 2. The algorithm starts with the largest sequence in *G. hipidula* and selects the best corresponding alignment in the *G. nigrocaulis* set as homologous pair. Subsequently the remaining set is iteratively processed and terminates when all sequences of one *Genlisea* species are assigned. The analysis resulted in 4,289 homologous intron and 5,395 homologous exon pairs.

To see whether intron size differences were caused by insertions in *G. hispidula* or deletions in *G. nigrocaulis*, intron sequences were checked for the presence of repetitive elements by BLASTN similarity searches against databases of Illumina reads derived from repeat clustering analysis. BLASTN e-value cutoff was set to 0.001 and minimal alignment length to 55 nucleotides. Such repeats (mostly MITEs or SINEs) were rare in both species (13 in *G. hispidula* and 6 in *G. nigrocaulis*) and not within homologous introns. We count this as indication that intron size differences were mostly caused by deletions in *G. nigrocaulis* instead of insertions in *G. hispidula*.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

5.2. Functional Annotation

'High confidence' and 'low confidence' genes were annotated using BLAST2GO (Conesa and Gotz 2008). To compare the annotation of the two species, we mapped all resulting gene ontology (GO) terms to the TAIR curated terms of Arabidopsis (Berardini *et al.* 2004). This mapping was done using a custom Scala script (http://www.scala-lang.org/). The set of current GO terms was downloaded from the Gene Ontology (http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ ext.obo) and incorporated into a Neo4j graph database. To map a GO term to a related term of interest, the graph was traversed (breadth-first) starting from the annotation to be mapped. If the traversal resulted in a path that ended in an interesting GO term, the candidate was mapped to this term. If multiple paths existed to the same GO term, it was counted only once. This procedure was applied for all pre-selected terms resulting in three ontology classes 1) biological processes, 2) molecular function and 3) cellular components. Tables S9-11 present for each of these three categories the proportion of genes related to a specific GO term (full list of GO annotation available in Supplementary Dataset S1). This comparative analysis shows for the majority of ontology classes a similar representation within the gene sets of both *Genlisea* species. Ontology terms with log2-fold >2 difference between species are labelled blue for excess in *G. hispidula*.

Table S9. Molecular Functions[a]

	G. hispidula		*G. nigrocaulis*		log_2-fold change
binding	18,212	20.74%	4,725	22.53%	-0,12
carbohydrate binding	147	0.17%	70	0.33%	-1,00
catalytic activity	6,569	7.48%	2,685	12.81%	-0,78
chromatin binding	436	0.50%	130	0.62%	-0,32
DNA binding	3,601	4.10%	1,113	5.31%	-0,37
enzyme regulator activity	260	0.30%	112	0.53%	-0,85

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

hydrolase activity	6,271	7.14%	2,378	11.34%	-0,67
kinase binding	52	0.06%	26	0.12%	-1,07
lipid binding	262	0.30%	121	0.58%	-0,95
motor activity	107	0.12%	36	0.17%	-0,49
nuclease activity	5,324	6.06%	297	1.42%	**2,10**
nucleic acid binding	13,259	15.10%	889	4.24%	1,83
nucleotide binding	4,920	5.60%	1,973	9.41%	-0,75
oxygen binding	44	0.05%	27	0.13%	-1,36
protein binding	1,856	2.11%	1,045	4.98%	-1,24
protein binding transcription factor activity	66	0.08%	29	0.14%	-0,88
receptor activity	113	0.13%	46	0.22%	-0,77
receptor binding	16	0.02%	11	0.05%	-1,53
RNA binding	8,870	10.10%	715	3.41%	1,57
structural molecule activity	507	0.58%	292	1.39%	-1,27
transferase activity	14,920	16.99%	3,245	15.48%	0,13
translation regulator activity	4	0.00%	3	0.01%	-1,65
transporter activity	2,013	2.29%	1,000	4.77%	-1,06
Total	**87,829**		**20,968**		

[a]Because individual genes may appear in more than one category, the total numbers may be higher than the actual gene counts per species

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Table S10. Biological Processes[a]

	G. hispidula		*G. nigrocaulis*		log_2-fold change
biosynthetic process	14,220	9.74%	3,882	7.33%	0,41
carbohydrate metabolic process	2,840	1.95%	1,492	2.82%	-0,53
catabolic process	2,609	1.79%	1,384	2.61%	-0,55
cell communication	322	0.22%	186	0.35%	-0,67
cell cycle	265	0.18%	121	0.23%	-0,33
cell death	85	0.06%	76	0.14%	-1,30
cell differentiation	378	0.26%	227	0.43%	-0,73
cell growth	681	0.47%	381	0.72%	-0,63
cellular amino acid metabolic process	1,130	0.77%	643	1.21%	-0,65
cellular component organization	5,150	3.53%	2,467	4.66%	-0,40
cellular homeostasis	238	0.16%	145	0.27%	-0,75
cellular process	34,761	23.81%	13,347	25.20%	-0,08
DNA metabolic process	20,898	14.32%	1,478	2.79%	**2,36**
electron transport chain	111	0.08%	59	0.11%	-0,55
embryo development	549	0.38%	307	0.58%	-0,62
flower development	87	0.06%	52	0.10%	-0,72
generation of precursor metabolites and energy	338	0.23%	192	0.36%	-0,65
growth	571	0.39%	315	0.59%	-0,60
leaf development	99	0.07%	55	0.10%	-0,61
lipid metabolic process	1,966	1.35%	1,163	2.20%	-0,71
metabolic process	30,225	20.71%	11,226	21.20%	-0,03
photosynthesis	89	0.06%	45	0.08%	-0,48
protein metabolic process	5,524	3.78%	2,047	3.87%	-0,03
protein modification process	2,794	1.91%	1,278	2.41%	-0,33
regulation of gene expression, epigenetic	641	0.44%	254	0.48%	-0,13
reproduction	40	0.03%	23	0.04%	-0,66
response to abiotic stimulus	3,236	2.22%	1,781	3.36%	-0,60
response to biotic stimulus	1,072	0.73%	586	1.11%	-0,59
response to endogenous stimulus	737	0.50%	436	0.82%	-0,71
response to external stimulus	410	0.28%	258	0.49%	-0,79
response to extracellular stimulus	244	0.17%	155	0.29%	-0,81
response to stress	5,335	3.65%	2,527	4.77%	-0,38
root development	151	0.10%	78	0.15%	-0,51
secondary metabolic process	403	0.28%	219	0.41%	-0,58
signal transduction	1,845	1.26%	999	1.89%	-0,58
translation	426	0.29%	236	0.45%	-0,61
transport	5,309	3.64%	2,752	5.20%	-0,51
tropism	199	0.14%	84	0.16%	-0,22
Total	**145,978**		**52,956**		

[a]Because individual genes may appear in more than one category, the total numbers may be higher than the actual gene counts per species

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Table S11. Cellular Components[a]

	G. hispidula		*G. nigrocaulis*		log_2-fold change
cell wall	381	1.45%	214	1.73%	-0,26
cytoplasm	1,329	5.05%	823	6.67%	-0,40
cytoskeleton	83	0.32%	39	0.32%	0,00
cytosol	1,407	5.34%	785	6.36%	-0,25
endoplasmic reticulum	403	1.53%	222	1.80%	-0,23
endosome	300	1.14%	163	1.32%	-0,21
extracellular matrix	6	0.02%	3	0.02%	-0,09
extracellular region	1,017	3.86%	601	4.87%	-0,33
extracellular space	10	0.04%	9	0.07%	-0,94
Golgi apparatus	797	3.03%	471	3.82%	-0,33
intracellular	1,059	4.02%	98	0.79%	**2,34**
membrane	2,932	11.13%	1,588	12.86%	-0,21
mitochondrion	3,199	12.15%	1,253	10.15%	0,26
nuclear membrane	13	0.05%	6	0.05%	0,02
nucleolus	232	0.88%	116	0.94%	-0,09
nucleoplasm	10	0.04%	6	0.05%	-0,36
nucleus	5,960	22.63%	2,787	22.58%	0,00
peroxisome	163	0.62%	93	0.75%	-0,28
plasma membrane	2,014	7.65%	1,131	9.16%	-0,26
plastid	4,226	16.05%	1,473	11.93%	0,43
ribosome	377	1.43%	193	1.56%	-0,13
thylakoid	142	0.54%	102	0.83%	-0,62
vacuole	277	1.05%	169	1.37%	-0,38
Total	**26,337**		**12,345**		

[a]Because individual genes may appear in more than one category, the total numbers may be higher than the actual gene counts per species

6. Polyploidy

6.1. Polyploidy identified by genome-wide *k*-mer statistics

The genome of *G. hispidula* is about 18-fold larger than that of *G. nigrocaulis* and presumably polyploid. To find out further evidence for this hypothesis, we analyzed the established gene sets by a *k*-mer index constructed with a cleaned set of WGS reads which were anchored to the reference genome. Regardless of the fragmented state of the *G. hispidula* genome assembly, this approach reveals multiple copies of a gene in polyploid genomes. We therefore analyzed the *k*-mer frequencies of genes with a detected orthologue in the other species and compared their respective

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

average *k*-mer frequency. In total, the coding DNA of 1,186 homologous gene pairs of the 'high confidence' gene set was compared. From the WGS data sets of *G. nigrocaulis* and *G. hispidula* we used those reads that align to the constructed reference to build a *k*-mer index using tallymer. The integrated reads represent the genome with a 350-fold coverage in *G. nigrocaulis* and 14-fold in *G. hispidula*. This index was applied to the respective data set of orthologous high confidence genes to compute their corresponding *k*-mer frequencies. The normalized *k*-mer frequencies support the assumption of mainly single copy genes in *G. nigrocaulis.* Only four genes of the analyzed set showed evidence of multiple copies. In contrast, the *G. hispidula* genome showed for the majority of analyzed genes (76.7%) multiple gene copies, indicated by average *k*-mer frequencies per gene of >1.5 (Supplementary Dataset S2). The complete set of analyzed genes showed an average copy number of 0.95 for *G. nigrocaulis*, and of 1.7 for *G. hispidula.* These values are close to a 1:2 ratio as expected in case of a WGD event. Because usually some genes may become lost after polyploidisation, an average gene copy number of 1.7 is compatible with tetraploidy in *G. hispidula*. We expect that the estimated values are computed with a minor underestimation, since *k*-mer analysis relies on a 100% identity in frequency estimation. For heterozygous positions or miss-assemblies respective *k*-mers have no or a lower representation in the computed index and consequently lead to a decrease of *k*-mer frequency. The effect is present for both *Genlisea* approximations and therefore both analyses are comparable.

6.2. Polyploidy identified by genome-wide SNP calling within transcripts

For both species an intra-specific transcript variant detection was performed for high confidence genes to detect heterozygous positions. As the first step, RNA-Seq reads from all tissues were aligned with ‘sensitive’ parameter settings by bowtie2

(Langmead and Salzberg 2012) to the set of 'high confidence' genes in both Genlisea species. PCR duplicates within the aligned read set were removed from the alignment by 'samtools rmdup'. The resulting alignments are affiliated into the variant calling process using VCFtools (Danecek *et al.* 2011). We removed putative false positive variants by discarding variants of inadequate read coverage (<10-fold) and insufficient variant quality (<150). The total number of SNPs was 19,391 for *G. nigrocaulis* and 140,556 for *G. hispidula*. Of the *G. nigrocaulis* SNPs 6,084 (31.38%) and of *G. hispidula* SNPs 91,420 (65.06%) were detected with an allele ratio (AR) of 0.4 to 0.6 (Fig. 3A). While in *G. nigrocaulis* these SNPs locate in 2,960 of the high confidence genes (~19.0%), in *G. hispidula* 15,562 'high confidence' genes (36.9%) indicate the existence of multiple gene copies. Assuming an equal heterozygosity level in both self-compatible species, 65.06% of alleles at heterozygous loci with a read frequency of about 1:1 (AR of 0.4-0.6,(23)) support the assumption of allotetraploidy in *G. hispidula*.

6.3. Copy number determination by genotype calling and haplotype phasing in a random gene set

Fifteen homologous gene pairs belonging to different *G. nigrocaulis* scaffolds were randomly selected for detecting their copy number in individuals of *G. nigrocaulis* and *G. hispidula* via genotype calling and haplotype phasing (Supplementary Dataset S3). In total 26 fragments (average 348 bp) of the selected genes carrying multiple heterozygous variant sites, as detected among transcripts, were amplified from three individuals of both species and analyzed by direct Sanger sequencing. Eleven genes (73.3%) in *G. hispidula* were found containing 63 heterozygous variants conserved between individual plants (SNPs or InDels). This variant type can be referred as interhomeolog or intergenomic polymorphism which is much more

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

frequent than intragenomic SNPs in polyploid plants (Trick *et al.* 2009). Two of the eleven genes have a single polymorphism in one fragment. Amplicons of 9 *G. hispidula* genes with multiple interhomeolog variants were cloned into the plasmid pJET1.2 (Thermo Scientific) and then Sanger sequenced for sorting variants into haplotypes (Fig. S6). For six genes two, and for three genes three different haplotypes of conserved heterozygous variants could be phased, confirming at least two copies of these genes in the *G. hispidula* genome.

7. Repetitive sequences

7.1. Clustering analysis

Repetitive elements were identified using similarity-based clustering of unassembled sequence reads (Novak *et al.* 2010) and further characterized using the REPEATEXPLORER pipeline (Novak *et al.* 2013). Clustering was performed with 860,000 and 7.2 million Ilumina reads for *G. nigrocaulis* and *G. hispidula*, respectively (Fig. S7). The classification of LTR retrotransposons into distinct lineages and clades was done using phylogenetic analysis of their reverse transcriptase (RT) sequences detected in contigs assembled for each repeat cluster (Novak, et al. 2013). Alignment of RT sequences was carried out with MUSCLE (EDGAR 2004) and the phylogenetic trees were calculated in SEAVIEW (Gouy *et al.* 2010) using the Neighbor-Joining algorithm with 1,000 bootstrap resamples. The trees were drawn and edited using FIGTREE (http://tree.bio.ed.ac.uk/software/figtree/, Fig. S8).

7.2. *K*-mer analysis of the *Genlisea* genome references

The genome reference sequences of *G. nigrocaulis* (~65 Mbp WGS assembly) and *G. hispidula* (~315 Mbp complete WGS assembly including all contigs >300 bp) were analyzed for their *k*-mer frequencies. The WGS reads (according Table S3) were

used to construct for both species an index of 21-mers. We applied this index to the reference sequence to compute *k*-mer frequencies using Tallymer (Kurtz, et al. 2008). The *k*-mer frequencies were normalized by the respective sequencing depth (350-fold for *G. nigrocaulis* and 14-fold for *G. hispidula*). The analysis revealed a very high proportion of unique *k*-mer sequences (83.4%) and a very low repetitive proportion within the *G. nigrocaulis* genome with an average normalized *k*-mer frequency of 7x (Table S12). Only 5.1% of all base pair positions of the genome assembly (~3.3 Mbp) show *k*-mer frequencies larger than 4x indicating few highly abundant sequences (Fig. S9). In comparison, the genome reference of *G. hispidula* with a 19x average *k*-mer frequency contains much lower unique *k*-mer sequences (36.9%) whereas the proportion of *k*-mers larger than 4x is much higher (>33.6% = 105.9 Mbp). When analyzing the genomes of different species for *k*-mer frequencies, the *k*-mer composition of a genome can also be compared as to the equal frequency of *k*-mer sequences. This comparative analysis of *k*-mer frequency between very closely related species could reveal which sequences multiplied, and which reduced abundance during evolution of the compared genomes. Therefore we looked for shared *k*-mers occurring with high frequency in the genomes of *G. nigrocaulis* and *G. hispidula* to detect such *k*-mers that became reduced or amplified in either species. To study this, we considered sequences as highly abundant, that have *k*-mer frequencies exceeding the average *k*-mer frequency by twofold. Thus, the respective threshold for *G. nigrocaulis* was set 14x and for *G. hispidula* 38x. In total, 183,658 individual 21-mers of the WGS assembly of *G. nigrocaulis* passed this criterion. In addition the *k*-mer index of *G. hispidula* was applied as well to the WGS assembly of *G. nigrocaulis.* This showed that only ~36% of individual *G.nigrocaulis k*-mers are shared with *G. hispidula*. Accordingly, the remaining ~64% of abundant *k*-mers are

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

not detected in the *G. hispidula* index, and the corresponding sequences could have either entered one genome after the split of the two *Genlisea* species or were removed from the other. It should be noted however that the performed *k*-mer comparison is valid only regarding gain or reduction of sequences which are identical between the compared species, but not for sequences that experienced nucleotide variation in one or both species. The distribution of *k*-mer frequency values in both *Genlisea* genomes is visualized in Fig. S10A, referring to the highly abundant *k*-mer sequences that originated from *G. nigrocaulis* and are shared with *G. hispidula*.

From the WGS assembly of *G. hispidula* 1.87 million individual 21-mers passed the frequency threshold of 38x. When applying the *k*-mer index of *G. nigrocaulis*, we identified only 1% of these *k*-mers to be shared between both *Genlisea* species. This suggests that the genome of *G. hispidula* contains many repetitive sequences that are not present in the *G. nigrocaulis* index. The reason for the large proportion of *k*-mer sequences not matching to the sister species is most likely the variability among these repetitive elements. To test this assumption, we performed a BLASTN analysis and reduced the 'percent identity' threshold to 95%. In contrast to the *k*-mer analysis that requires 100% identity, the BLAST analysis revealed ~56% of the analyzed highly abundant sequences of *G. hispidula* to be present in the *G. nigrocaulis* genome, although with much lower abundance and often affected by mutations. The distribution of frequencies between the shared and highly abundant *k*-mer sequences originating from the *G. hispidula* genome is illustrated in Fig. S10B. The comparison of both distributions indicates that a considerable amount of *k*-mers are abundant in *G. nigrocaulis* and occur with lower frequency in *G. hispidula* (Fig. S10A, hexagons above medial diagonal). In addition there is a substantial quantity of *k*-mers frequent in *G. nigrocaulis* that has even a higher frequency in *G. hispidula*

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

(hexagons under medial diagonal). For *k*-mers classified as highly frequent in *G. hispidula* (Fig. S10B) the proportion of repeat elements which are of substantially lower frequency in *G. nigrocaulis* is very high (hexagons in light blue cluster at the bottom of Fig. S10B). We propose that these *k*-mers with highly dissimilar repetitiveness belong to repeat elements that are highly amplified in *G. hispidula*.

7.3. SINEs in *Genlisea*

To reveal the presence of short interspersed nuclear elements (SINE) in *Genlisea*, we used the SINE-Finder tool (Wenke *et al.* 2011) and performed a screening in the WGS assembly of *G. nigrocaulis* and *G. hispidula*. These short (<500bp) and nonautonomous retroelements usually terminate at the 3' end with nucleotide stretches of poly(A) or poly(T), or other simple sequence motifs. Flanking target site duplications (TSDs) define the composite structure of SINEs (Wenke, et al. 2011). Our analysis proofed the existence of SINEs in both *Genlisea* species (282 elements in *G. nigrocaulis* and 1,481 elements in *G. hispidula*). The cumulative length covers 81.5 kbp (<0.1%) in *G. nigrocaulis* and 352.7 kbp in *G. hispidula*, while the medium size of SINE is 289 bp and 238 bp, respectively. For *G. hispidula* this is likely an underestimation due to the fragmentation of the WGS assembly and coinciding repeat elements.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Table S12. *K*-mer statistics of *Genlisea* genomes

	G. nigrocaulis			*G. hispidula*		
k-mer frequency	#position	percentage of WGS assembly	∑ WGS assembly representation	#position	percentage of WGS assembly	∑ WGS assembly representation
0	5,984,071	9.22	9.22	9,040,785	2.87	2.87
1	54,162,167	83.42	92.64	116,103,720	36.92	39.79
2	1,454,167	2.24	94.87	83,469,489	26.54	66.33
4	1,531,525	2.36	97.23	49,150,367	15.63	81.96
8	534,770	0.82	98.06	20,913,558	6.65	88.61
16	352,393	0.54	98.60	13,058,398	4.15	92.76
32	375,428	0.58	99.18	8,234,696	2.62	95.38
64	232,803	0.36	99.54	5,356,894	1.70	97.08
128	106,954	0.16	99.70	3,637,012	1.16	98.24
256	31,145	0.05	99.75	2,410,706	0.77	99.01
512	9,803	0.02	99.76	1,481,682	0.47	99.48
1,024	30,812	0.05	99.81	805,276	0.26	99.73
2,048	32,972	0.05	99.86	482,365	0.15	99.89
4,096	76,308	0.12	99.98	208,145	0.07	99.95
8,192	12,745	0.02	100.00	93,208	0.03	99.98
>8,192	0	0.00	100.00	55,780	0.02	100.00

7.4. Mobile elements in RNA-seq data

To study the proportion of mobile elements that are transcribed in both Genlisea species, we screened our Blast2GO annotation (Supplementary section 5.2.) for 'transposase'-like features. In *G. hispidula* 6,204 gene models and in *G. nigrocaulis* 352 gene models were filtered because of their high similarity to mobile elements. Among these, several elements are not categorized in more detail ('unclassified'), but the majority of described elements of *G. hispidula* (875 and 2,724) and *G. nigrocaulis* (66 and 95) are annotated as 'Ty1/copia' or 'Ty3/gypsy' retroelements. Subsequently, for both species RNA-seq derived reads (all tissues) were aligned to all transposase gene models by bowtie2. In both Genlisea species transcripts of these elements were detected. With 3,349 transcribed transposase-like genes, mobile elements seem to be more active in *G. hispidula* than in *G. nigrocaulis* with 207 transposase-like genes aligned to an RNA-seq read. Remarkably, only five of

the *G. nigrocaulis* genes capture 60% of the aligned RNA-seq reads. Three of the five represent Ty1/copia elements as revealed by the Blast2GO annotation. The remaining two are 'unclassified' mobile elements. The retroelement of the *Bianca* lineage is the most strongly transcribed element of the *G. nigrocaulis* genome. The abundance of the *Bianca* element in the *G. nigrocaulis* genome of becomes obvious in Fig. S11 constructed according to Darzentas (Darzentas 2010) and illustrating the 23 scaffolds of *G. nigrocaulis* with the highest abundance of *Bianca*-related sequences.

Table S13. Transposase genes of mobile elements transcribed in *Genlisea* species.

Species	*G. hispidula*		*G. nigrocaulis*	
Gene models class	HC	LC	HC	LC
# gene models annotated as 'retroelement'	2,491	3,713	219	133
# gene models mapped by RNA-seq reads	2,321	1,028	181	26
Total mapped RNA-seq reads	400,733	12,135	234,538	378
# gene models with >100 mapped reads	454	10	78	1
# gene models with >500 mapped reads	167	2	41	0
# gene models capturing 70% of all reads	130	41	10	4
# gene models capturing 60% of all reads	84	12	5	3

8. Ks-based dating and genome size evolution within *Genlisea*

Fifty homologous gene pairs from both species with two homoeologous copies in *G. hispidula* were randomly selected from the high-confidence gene sets. They were aligned with MUSCLE (EDGAR 2004) and alignments were checked by eye and, where necessary, corrected manually. Reading frames were inferred, introns removed, the exonic sequences of each locus degapped (Supplementary Dataset S4), and pairwise Ks values were calculated for the *G. nigrocaulis* – *G. hispidula* orthologs as well as for the two homoeologous copies of *G. hispidula* with KAKS_CALCULATOR

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

(Zhang *et al.* 2006) using Model Averaging (MA). To arrive at approximate age estimates, the median values for the orthologous and homoeologous Ks values (Fig. S12) were calculated and converted assuming a neutral mutation rate of 1.5×10^{-9} mutations per site per year with the formula

$$Age = Ks / 2 \times 1.5 \times 10^{-9}.$$

This resulted in an age estimate for the split between the lineages leading to *G. nigrocaulis* and *G. hispidula* of 29 My (Fig. 1). This number is much higher than the 13 My calculated by Ibarra-Laclette et al. (Ibarra-Laclette *et al.* 2013) based on chloroplast *trn*L-F sequences (see their supplementary Figure 39). The lack of fossils in Lentibulariaceae poses severe problems for calibration of phylogenetic trees. Moreover, Ibarra-Laclette et al. (Ibarra-Laclette, et al. 2013) report doubled mutation rates at the chloroplast *trn*L-F locus for the lineage leading to section *Genlisea* (represented by *G. aurea* and *G. guianensis*) in comparison to the lineage of *G. hispidula* (see their supplementary Figure 40). As it cannot be said whether elevated mutation rates of one chloroplast site might be indicative of generally higher mutation rates also for the coding regions of the 50 selected genes from the nucleus, we here assume that the split between the lineages of *G. hispidula* and *G. nigrocaulis* might date back to a time between Upper Oligocene to Middle Miocene, i.e. between 29 and 13 My ago. Age calculations for the split between both parental lineages contributing to allotetraploid *G. hispidula* based on the *G. hispidula* homoeologues at the 50 compared nuclear loci resulted in 20 My. This provides a maximum age for the allopolyploidisation event, as hybridization must have happened after the two lineages were already separated. Ibarra-Laclette et al. ((Ibarra-Laclette, et al. 2013), supplementary Figure 40) found no evidence for elevated mutation rates in the *G. hispidula* lineage, which means that they might be closer to the Brassicaceae rates

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

assumed here. A minimum age cannot be provided, as the homoeologous loci might have diversified before (i.e. within the parental lineages) as well as after allotetraploidisation (i.e. within the merged genomes of the *G. hispidula* lineage).

When taking into account that the genome of *G. hispidula* is tetraploid, its diploid progenitors should have had genomes of ~800 Mbp (1C), similar to the value found in *G. africana* (750 Mbp; Fig. 1) that belongs to the diploid sister group of *G. hispidula*/*G. subglabra*. Taking also into account that the size differences between the *G. hispidula* and *G. nigrocaulis* lineages are, besides WGD, mostly due to DNA loss in the latter, we face a loss of 714 Mbp (difference between 800 Mbp in the diploid *G. hispidula* progenitor and 86 Mbp in *G. nigrocaulis*) within either 29 or 13 My. Assuming a generation time of 2 years (both lineages consist of annual and perennial species; Fig. 1) this means that on average 50–110 bp of DNA were jettisoned from the 1C genome every generation within the lineage leading to *G. nigrocaulis*, which correspond to 0.02–0.03% DNA loss per generation.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.

Bennett, M.D., Leitch, I.J., Price, H.J. and Johnston, J.S. (2003) Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Ann. Bot.*, **91**, 547-557.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D. and Rhee, S.Y. (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol.*, **135**, 745-755.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W. (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578-579.

Boetzer, M. and Pirovano, W. (2012) Toward almost closed genomes with GapFiller. *Genome Biol.*, **13**, R56.

Clarke, J.D. (2009) Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harb. Protoc.*, **2009**, pdb prot5177.

Conesa, A. and Gotz, S. (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*, **2008**, 619832.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. and Durbin, R. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156-2158.

Darzentas, N. (2010) Circoletto: visualizing sequence similarity with Circos. *Bioinformatics*, **26**, 2620-2621.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.

Dolezel, J., Greilhuber, J. and Suda, J. (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.*, **2**, 2233-2244.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Fleischmann, A. (2012) *Monograph of the Genus Genlisea* Poole, Dorset, England: Redfern Natural History Productions.

Galbraith, D.W., Harkins, K.R., Maddox, J.M., Ayres, N.M., Sharma, D.P. and Firoozabady, E. (1983) Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science*, **220**, 1049-1051.

Gouy, M., Guindon, S. and Gascuel, O. (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221-224.

Ibarra-Laclette, E., Lyons, E., Hernandez-Guzman, G., Perez-Torres, C.A., Carretero-Paulet, L., Chang, T.H., Lan, T., Welch, A.J., Juarez, M.J., Simpson, J., Fernandez-Cortes, A., Arteaga-Vazquez, M., Gongora-Castillo, E., Acevedo-Hernandez, G., Schuster, S.C., Himmelbauer, H., Minoche, A.E., Xu, S., Lynch, M., Oropeza-Aburto, A., Cervantes-Perez,

S.A., de Jesus Ortega-Estrada, M., Cervantes-Luevano, J.I., Michael, T.P., Mockler, T., Bryant, D., Herrera-Estrella, A., Albert, V.A. and Herrera-Estrella, L. (2013) Architecture and evolution of a minute plant genome. *Nature*, **498**, 94-98.

Jasencakova, Z., Soppe, W.J., Meister, A., Gernand, D., Turner, B.M. and Schubert, I. (2003) Histone modifications in Arabidopsis- high methylation of H3 lysine 9 is dispensable for constitutive heterochromatin. *The Plant journal : for cell and molecular biology*, **33**, 471-480.

Kelley, D.R., Schatz, M.C. and Salzberg, S.L. (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, **11**, R116.

Kersey, P.J., Staines, D.M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J.C., Hughes, D.S., Keenan, S., Kerhornou, A., Koscielny, G., Langridge, N., McDowall, M.D., Megy, K., Maheswari, U., Nuhn, M., Paulini, M., Pedro, H., Toneva, I., Wilson, D., Yates, A. and Birney, E. (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91-97.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Kurtz, S., Narechania, A., Stein, J.C. and Ware, D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, **9**, 517.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357-359.

Li, L., Stoeckert, C.J., Jr. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178-2189.

Lysak, M., Fransz, P. and Schubert, I. (2006) Cytogenetic analyses of Arabidopsis. *Methods Mol. Biol.*, **323**, 173-186.

Marcais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764-770.

Novak, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792-793.

Prlic, A., Yates, A., Bliven, S.E., Rose, P.W., Jacobsen, J., Troshin, P.V., Chapman, M., Gao, J., Koh, C.H., Foisy, S., Holland, R., Rimsa, G., Heuer, M.L., Brandstatter-Muller, H., Bourne, P.E. and Willis, S. (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**, 2693-2695.

Schmidt-Lebuhn, A.N., Fuchs, J., Hertel, D., Hirsch, H., Toivonen, J. and Kessler, M. (2010) An Andean radiation: polyploidy in the tree genus Polylepis (Rosaceae, Sanguisorbeae). *Plant Biol. (Stuttg.)*, **12**, 917-926.

Stanke, M. and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, **33**, W465-W467.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Swain, M.T., Tsai, I.J., Assefa, S.A., Newbold, C., Berriman, M. and Otto, T.D. (2012) A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat. Protoc.*, **7**, 1260-1284.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511-515.

Trick, M., Long, Y., Meng, J. and Bancroft, I. (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant Biotechnol. J.*, **7**, 334-346.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*, **23**, 3117-3128.

Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K. and Yu, J. (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics*, **4**, 259-263.

Supplementary Data Set

Supplementary Dataset S1: BLAST2GO tables of functional annotation of 'high confidence' and 'low confidence' genes (Excel)

Supplementary Dataset S2: Average *k*-mer frequency of 1,186 homologous genes supporting polyploidy in *G. hispidula* (Excel)

Supplementary Dataset S3: Genotype calling and haplotype phasing in a set of randomly selected gene pairs of *G. nigrocaulis* and *G. hispidula.* (Excel)

Supplementary Dataset S4: Alignment of coding sequences of 50 randomly selected genes used for Ks-based age estimations.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Supplemental Figures

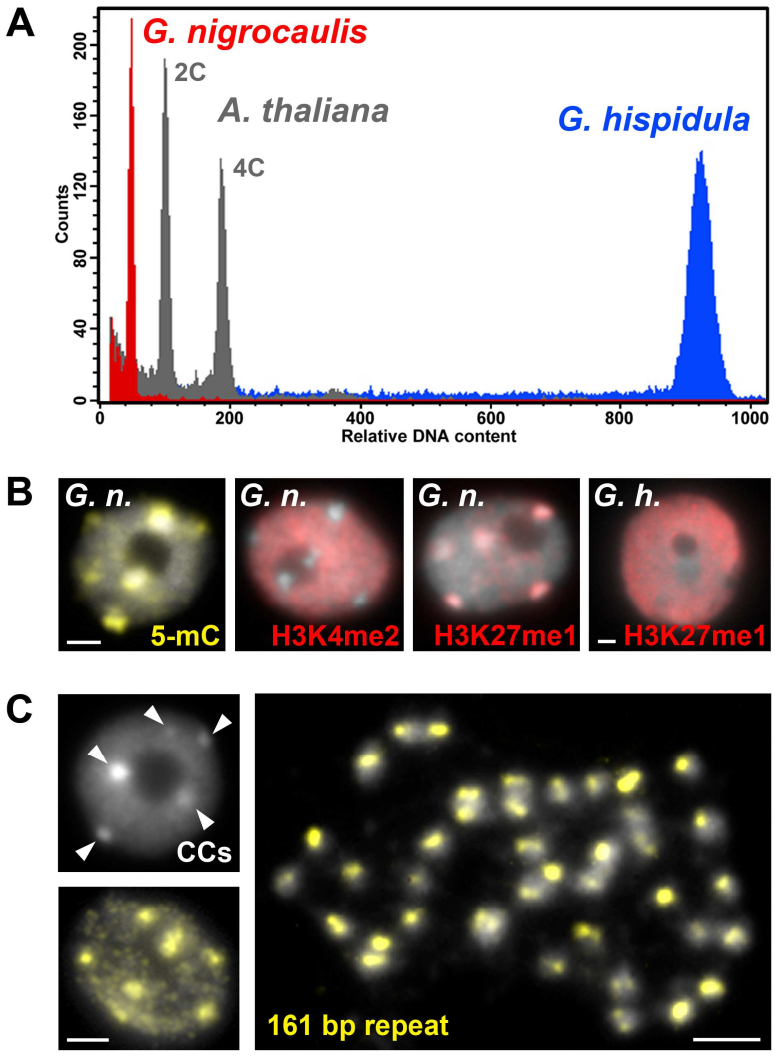


Figure S1. DNA content and heterochromatin in *G. nigrocaulis*. (**A**) Histogram of nuclear DNA contents with *A. thaliana* as reference standard. For better visualization of the size difference, separate histograms were superimposed. For genome size measurements internal reference standards were used. (**B**) Interphase nuclei after DNA methylation (left), euchromatin-specific H3K4me2 labeling (middle 1), heterochromatin-specific H3K27me1 labeling (middle 2), and for comparison an interphase nucleus of *G. hispidula* with a rather uniform distribution of H3K27me1 signals (right). **C,** *G. nigrocaulis* heterochromatic chromocenters in interphase nuclei (left, top) are mainly composed of the single highly abundant 161 bp tandem repeat (left, bottom), which is present on each of the 2n=40 chromosomes (right), likely at centromeric position, and is a unique or low copy sequence in *G. hispidula*. Each bar represents 2.5 µm.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

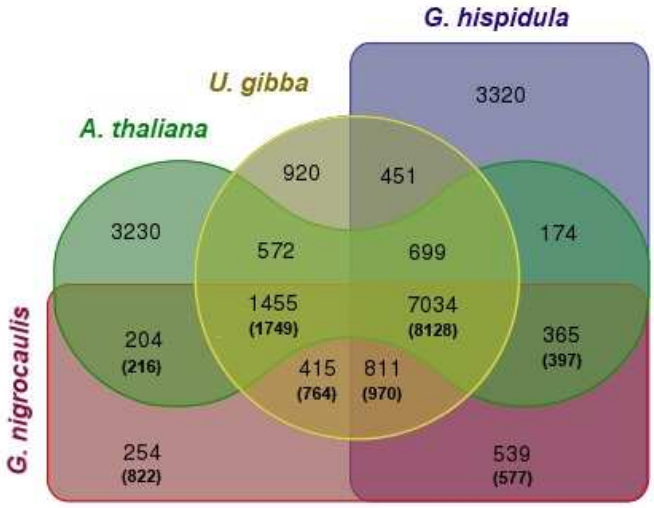


Figure S2. Venn diagram summarizing the distribution of orthologous gene cluster shared between *G. nigrocaulis*, *G. hispidula*, *U. gibba* and *A. thaliana*. The number of genes in *G. nigrocaulis* for the corresponding orthologous group is given in brackets.

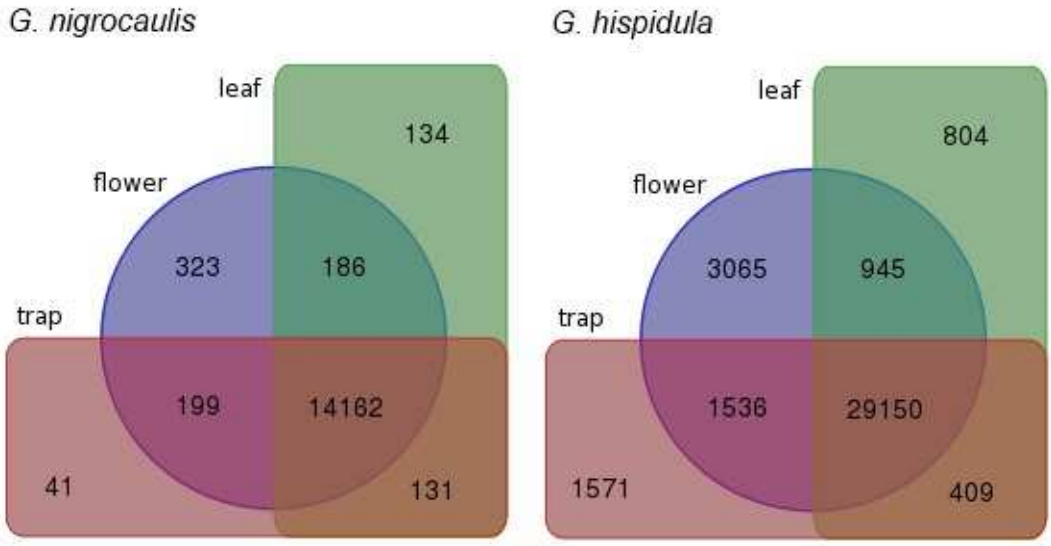


Figure S3. Expression analysis of high confidence genes in trap, leaf and flower of both *Genlisea* species.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

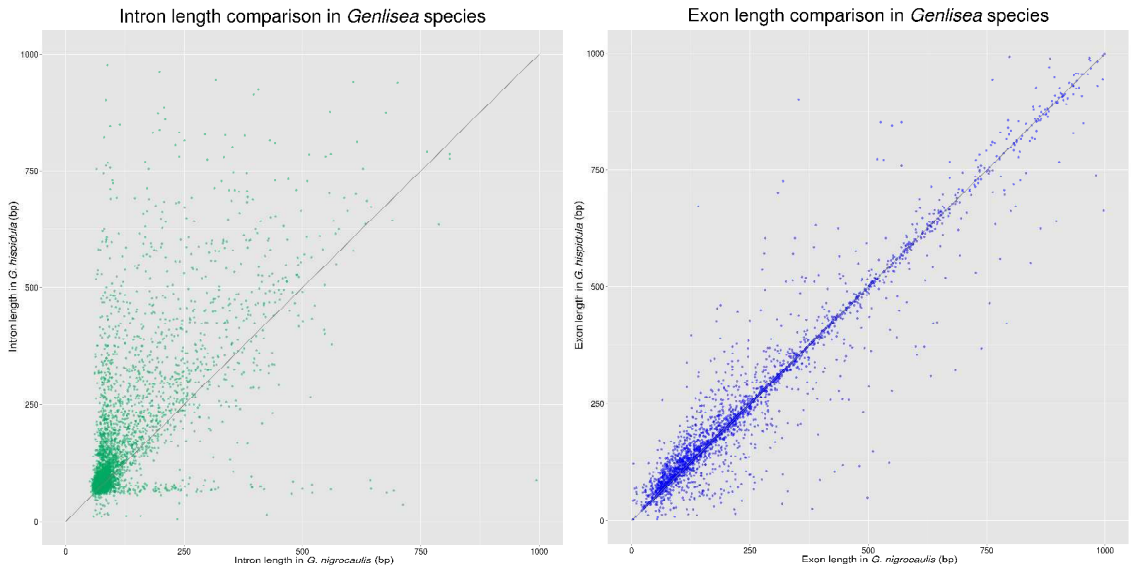
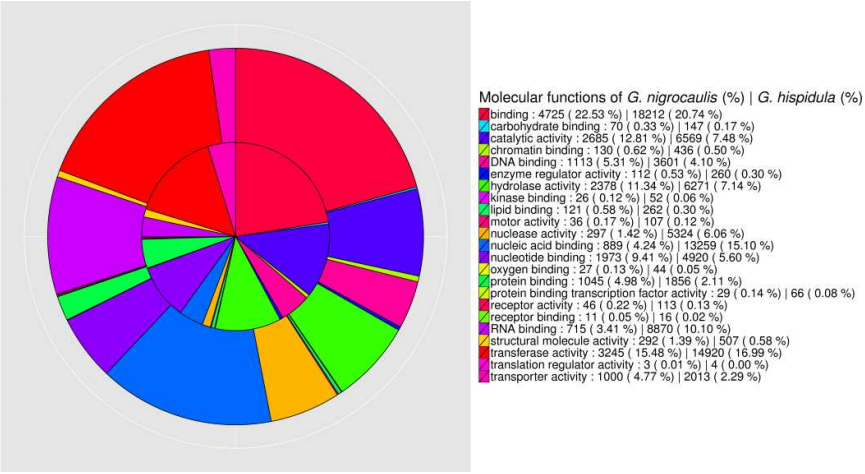
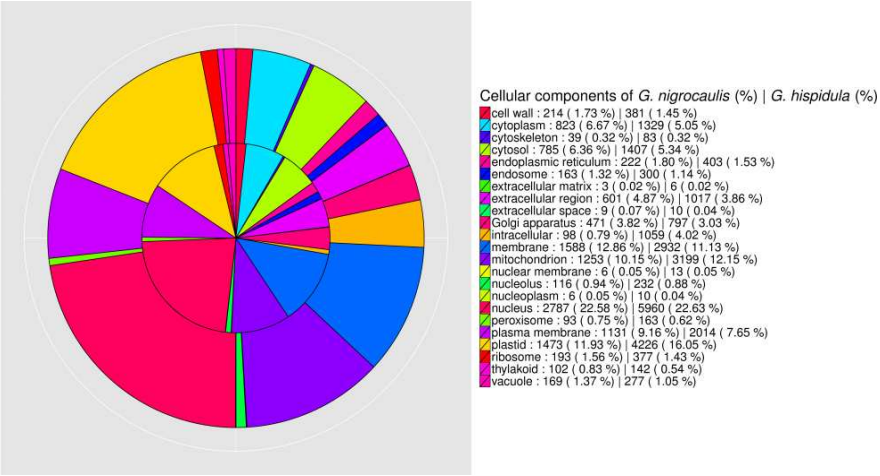


Figure S4. Complete intron and exon size distribution compared between 1,186 homologous genes of *G. hispidula* and *G. nigrocaulis*. Intron length comparison utilized a subset of 814 gene pairs with at least one intron sequence in both species.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021



(A) Molecular functions



(B) Cellular components

Figure S5. Gene ontology annotation of of *G. nigrocaulis* and *G. hispidula.* (**A**) Molecular functions; (**B**) Cellular components. Outer circle: *G. hispidula*, inner circle: *G. nigrocaulis.*

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

Figure S6. Examples of genotype calling and haplotype phasing in conserved regions of *G. hispidula* homoeologous loci. Conserved heterozygous variants are represented by differently colored rings corresponding to nucleotide variants of the SNPs or InDels. Interhomeolog polymorphism in polyploid species undetectable in consensus NGS contigs can be phased into haplotypes by cloning and Sanger sequencing. Numbers in clonal haplotypes indicate the number of supporting plasmid sequences. (*) Additional allelic variants revealed 4 haplotypes, as possible in tetraploid genomes.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

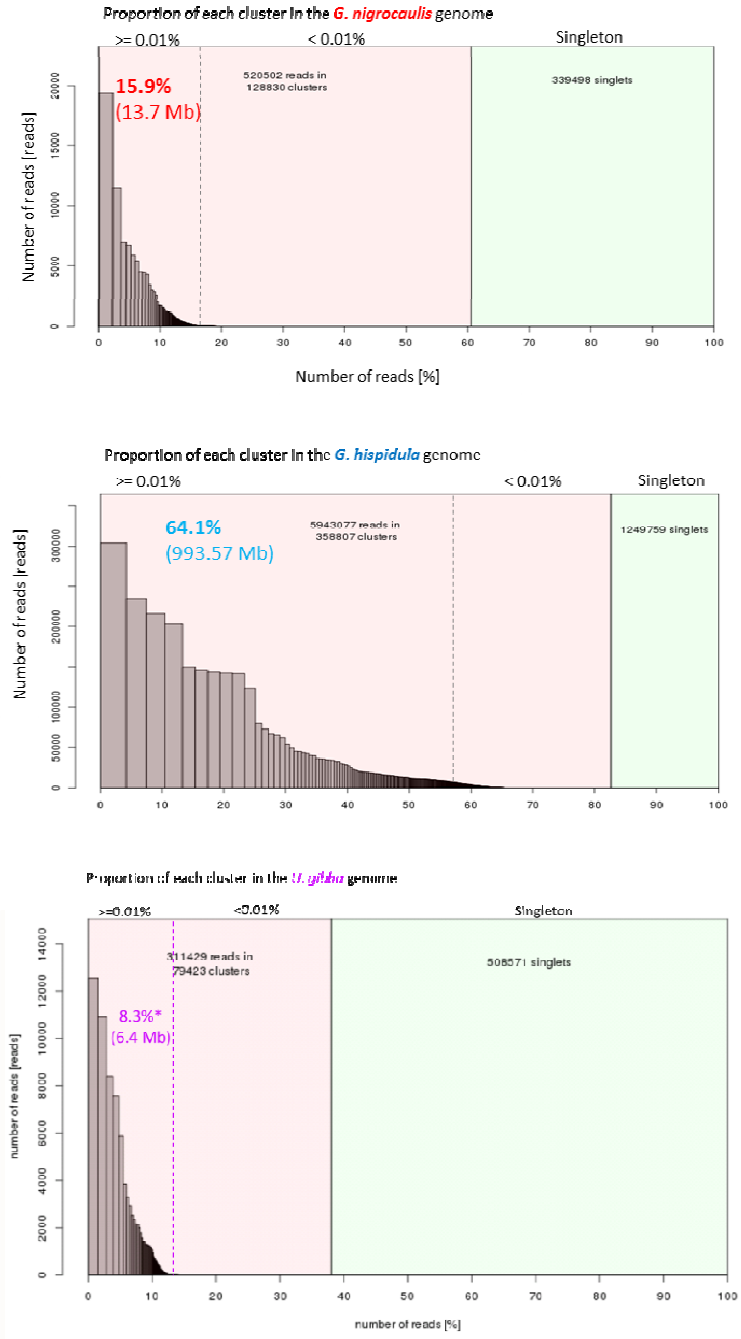
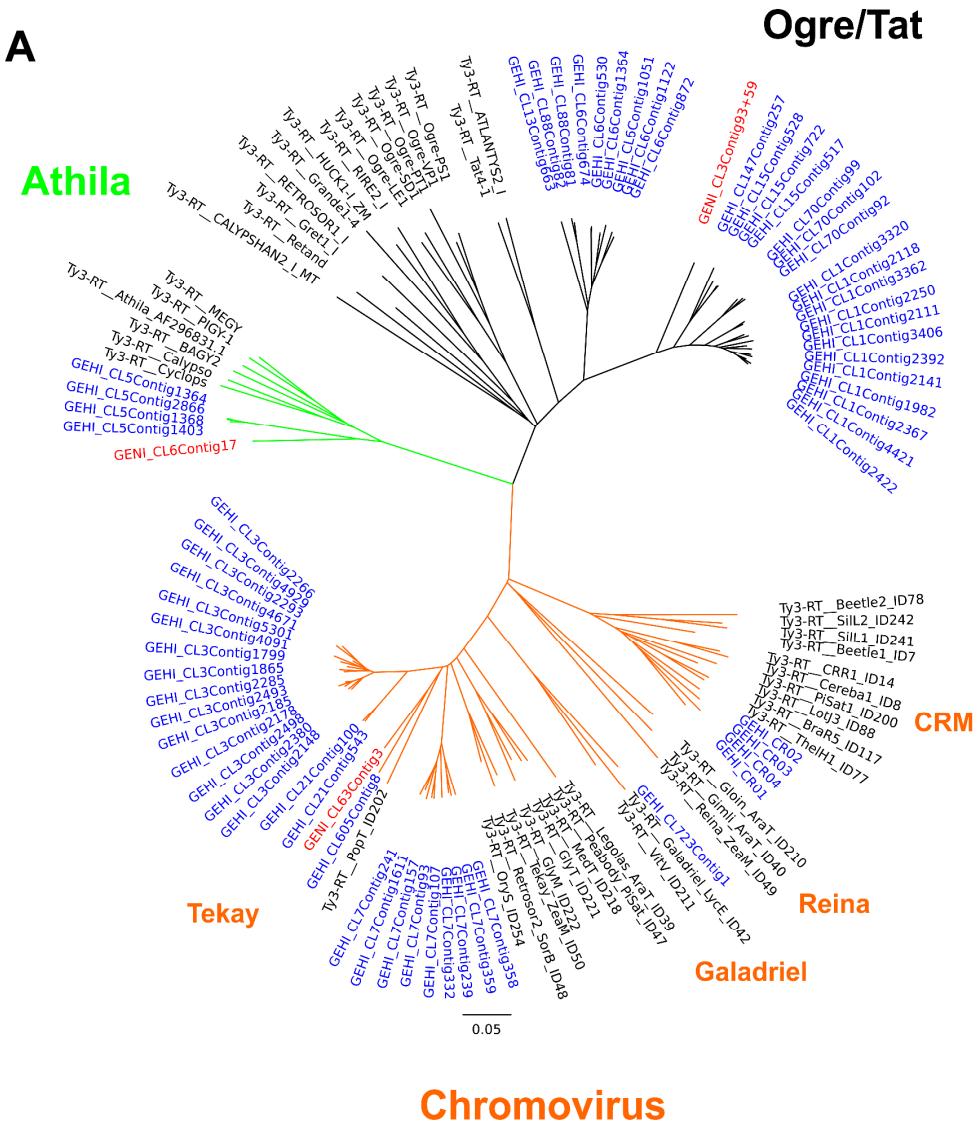


Figure S7. Cluster analysis of randomly selected WGS sequence reads. 15.9% (13.7 Mbp/1C) of *G. nigrocaulis,* 64.1% (993.5 Mbp/1C) of *G. hispidula,* and 8.3% (6.4 Mbp/1C) of *Utricularia gibba* genomes were identified as repetitive sequences (each >=0.01% of the genome). *The value for *U. gibba* is based on published sequence reads (Ibarra-Laclette, et al. 2013), after subtraction of 3.9% of viral and bacterial contamination.


A
Ogre/Tat
Athila
CRM
Reina
Galadriel
Tekay
Chromovirus
0.05

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

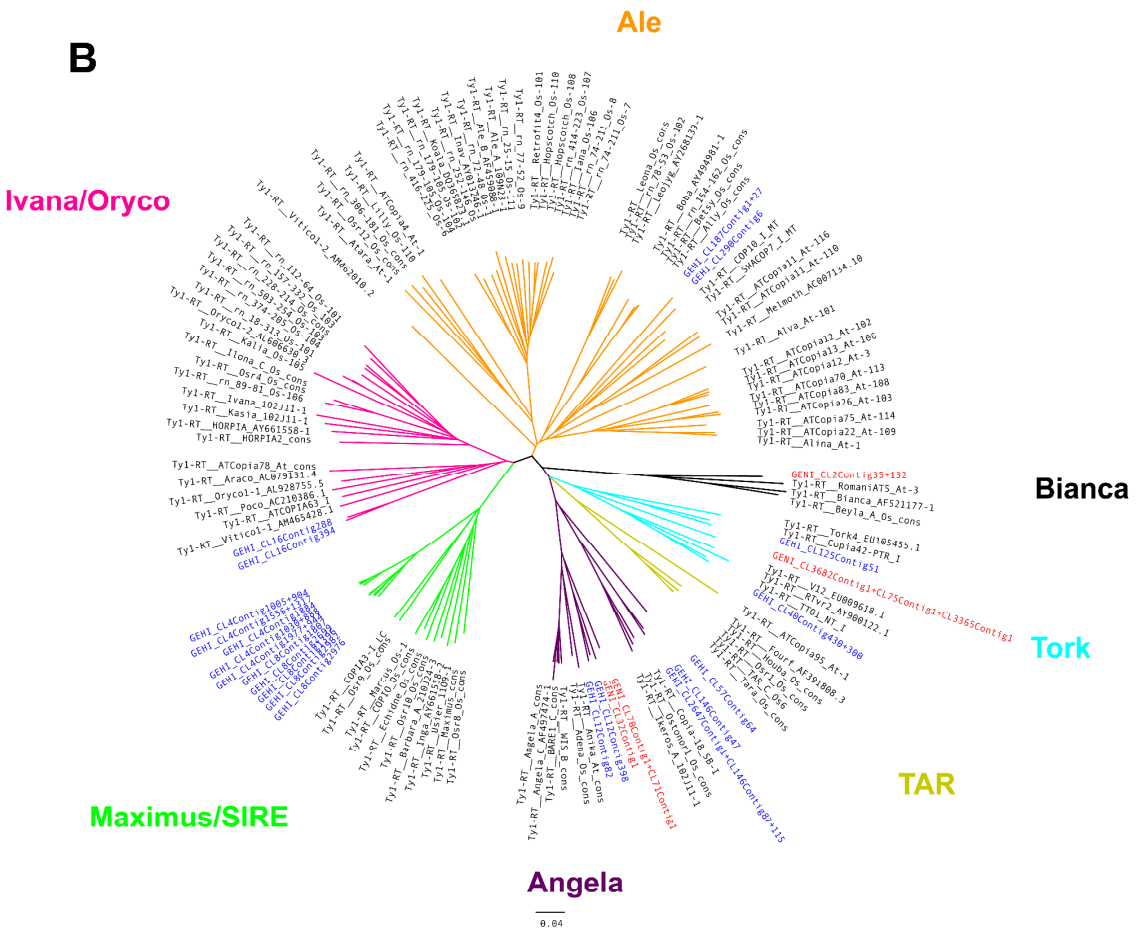


Figure S8. Unrooted neighbor-joining trees inferred from a comparison of reverse transcriptase (RT) domain sequences of LTR-retrotransposons identified in the read clustering analysis. The analysis demonstrates that Ty3/gypsy (A) and Ty1/copia (B) retrotransposons in *G. hispidula* are not only more abundant than in *G. nigrocaulis* (as shown in Fig. 2) but also more divergent in their sequences. RT sequences from *G. nigrocaulis* and *G. hispidula* are shown in red and blue, respectively. RT sequences from the representatives of major retrotransposon lineages are in black.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

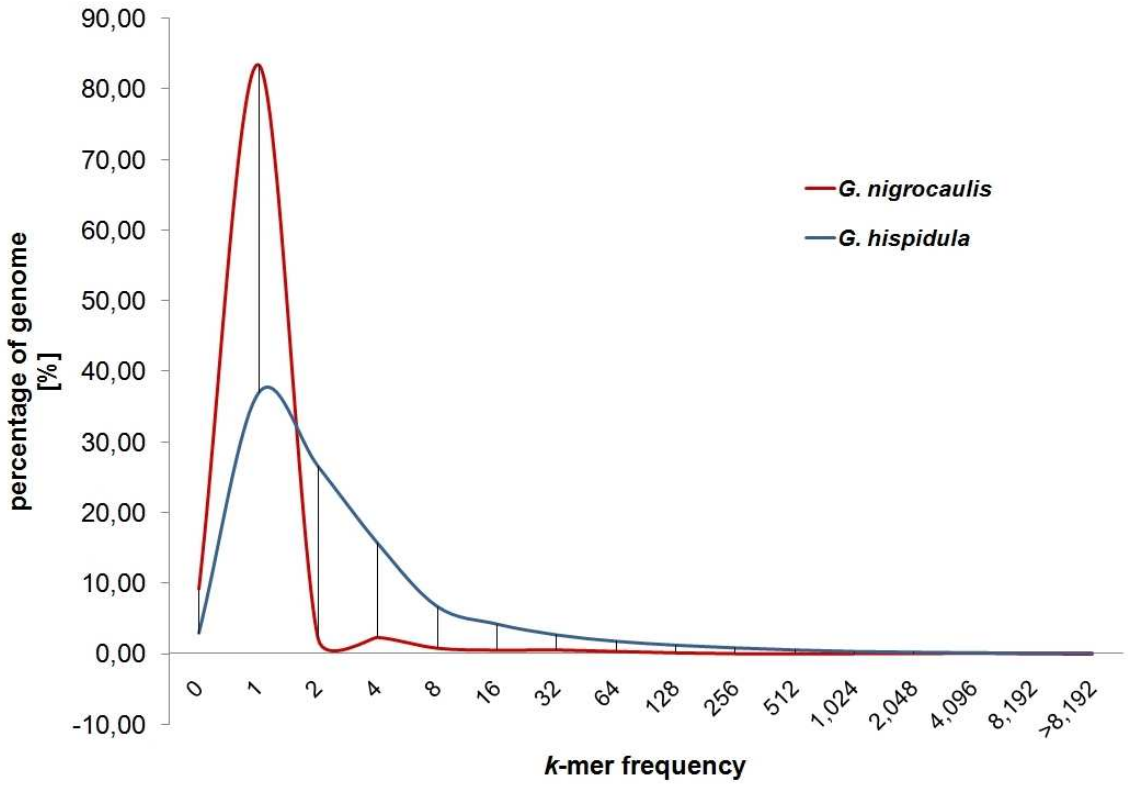


Figure S9. *K*-mer composition of the *Genlisea* genomes. For the WGS assembly of *G. nigrocaulis* (blue) a high proportion of unique and low-copy sequences is detected (83.4% of WGS assembly positions are represented by unique sequences). In contrast, only 36.9% of the sequence position in the WGS assembly of *G. hispidula* is estimated to be unique (red).

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

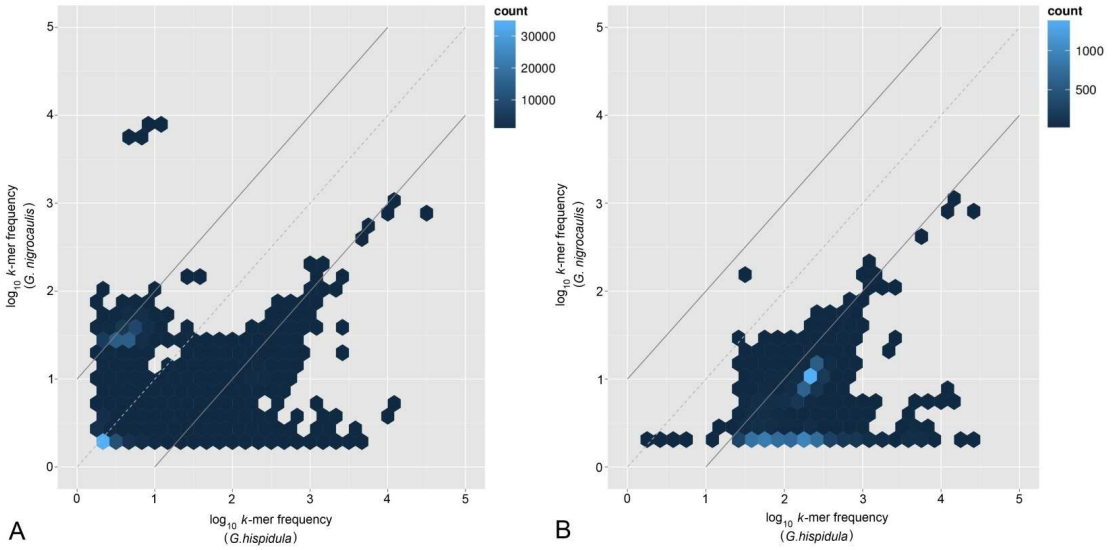


Figure S10. **Comparison of abundant *k*-mer sequences between both *Genlisea* species**. (**A**) Abundant *k*-mer sequences detected in the WGS assembly of *G. nigrocaulis* and corresponding *k*-mer frequencies observed in both *Genlisea* species. (**B**) Abundant *k*-mer sequences detected in the WGS assembly of *G. hispidula* and corresponding *k*-mer frequencies observed in both *Genlisea* species.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

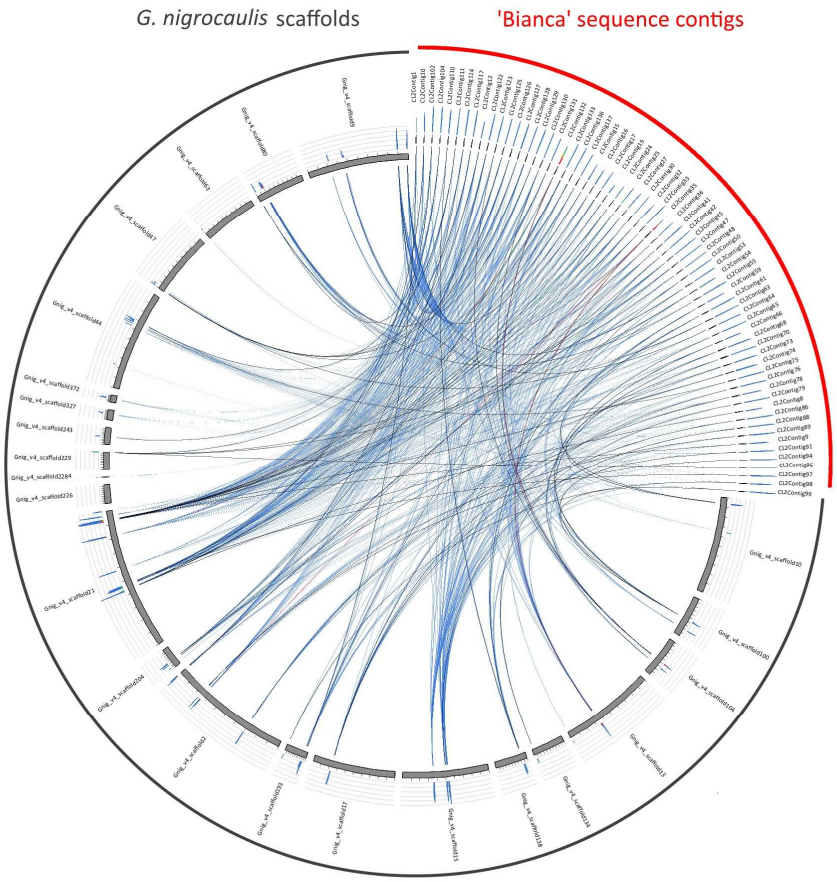


Figure S11. Twenty three genomic scaffolds of *G. nigrocaulis* that harbor the most *Bianca* retroelement sequences.

The Plant Genome Accepted paper, posted 09/02/2015. doi:10.3835/plantgenome2015.04.0021

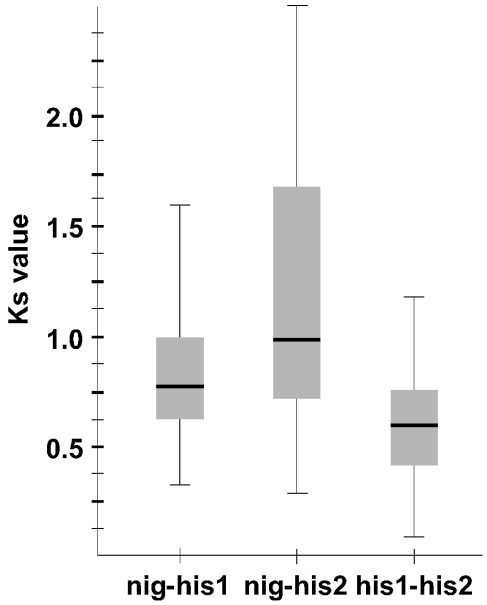


Figure S12. Box-and-whisker plots for the distribution of pairwise Ks values of 50 randomly selected nuclear genes of *G. nigrocaulis* (diploid) and *G. hispidula* (tetraploid, therefore two copies are present). Black bars indicate median values.