scientific reports

OPEN

Unraveling the maternal and paternal origins of allotetraploid *Vigna reflexo-pilosa*

Jayern Lee[1,6], Yang Jae Kang[2,4], Halim Park[2], Sangrea Shim[3], Jungmin Ha[5], Taeyoung Lee[6], Moon Young Kim[1,7] & Suk-Ha Lee[1,7]

The genomic structures of *Vigna hirtella* Ridl. and *Vigna trinervia* (B.Heyne ex Wight & Arn.) Tateishi & Maxted, key ancestral species of the allotetraploid *Vigna reflexo-pilosa* var. *glabra* (Roxb.) N.Tomooka & Maxted, remain poorly understood. This study presents a comprehensive genomic comparison of these species to deepen our knowledge of their evolutionary trajectories. By comparing the genomic profiles of *V. hirtella* and *V. trinervia* with those of *V. reflexo-pilosa*, we investigate the complex genomic mechanisms underlying allopolyploid evolution within the genus *Vigna*. Comparison of the chloroplast genome revealed that *V. trinervia* is closely related to *V. reflexo-pilosa*. De novo assembly of the whole genome, followed by synteny analysis and Ks value calculations, confirms that *V. trinervia* is closely related to the A genome of *V. reflexo-pilosa*, and *V. hirtella* to its B genome. Furthermore, the comparative analyses reveal that *V. reflexo-pilosa* retains residual signatures of a previous polyploidization event, particularly evident in higher gene family copy numbers. Our research provides genomic evidence for polyploidization within the genus *Vigna* and identifies potential donor species of allotetraploid species using de novo assembly techniques. Given the Southeast Asian distribution of both *V. hirtella* and *V. trinervia*, natural hybridization between these species, with *V. trinervia* as the maternal ancestor and *V. hirtella* as the paternal donor, seems plausible.

Polyploidy is a remarkable biological phenomenon characterized by the presence of more than two sets of chromosomes in an organism[1]. Among the different types of polyploidy, allotetraploidy occurs when two different genomes combine as a result of hybridization between different species or varieties. The effects of allotetraploidy on plants include changes in gene expression, increased size and growth rate, and altered reproductive behavior[2,3]. As a result, it has been the focus of extensive research in the plant sciences[4,5].

Soybean (*Glycine max*), an important legume crop, exhibits allotetraploidy and has been extensively studied due to its agricultural importance[6]. The availability of a reference genome for soybean[7] has facilitated these studies; however, understanding the genetic components provided by the ancestral species remains elusive[8]. As a result, the distribution and evolutionary history of donor genomic components in allotetraploid soybean are not fully understood.

Following polyploidization, diploidization and fractionation mechanisms are expected to mitigate the potentially deleterious effects of increased gene dosage on plant adaptability[9]. Diploidization involves halving the chromosomal complement of the polyploid genome, resulting in a diploid-like genome capable of restoring regular meiotic processes and sexual reproduction[10]. Fractionation, on the other hand, refers to the selective loss of redundant or nonessential genes following polyploidization[11]. This process reduces gene dosage and mitigates potential genetic imbalances caused by gene duplication[11]. Through the rationalization of their genetic makeup, plants have the capacity to shape their own genome architecture, influencing the emergence

[1]Department of Agriculture, Forestry and Bioresources and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea. [2]Division of Bio & Medical Bigdata Department (BK4 Program), Gyeongsang National University, Jinju, Republic of Korea. [3]Department of Forest Resources, College of Forest and Environmental Sciences, Kangwon National University, Chuncheon, Republic of Korea. [4]Division of Life Science Department at, Gyeongsang National University, Jinju, Republic of Korea. [5]Department of Plant Science, Gangneung-Wonju National University, Gangneung, Republic of Korea. [6]Macrogen, Seoul, Republic of Korea· [7]Plant Genomics and Breeding Institute, Seoul National University, Seoul, Republic of Korea. email: sukhalee@snu.ac.kr

of novel gene functionalities, regulatory networks, and phenotypic traits during evolution[12]. This type of large scaled chromosome rearrangement and rebuilding has been documented in previous studies involving *Brassica napus*[13,14], *Tragopogon* allopolyploids[15], and *Pyrus bretschneideri*[16].

To gain valuable insights into the gene-level consequences of diploidization and fractionation processes in the context of known donor and allotetraploid species, it is imperative to study genetic interactions in species with known genetic backgrounds.

The *Vigna*, a genus within the legume family, encompasses over 100 plant species, with agronomic importance attributed to certain key species such as cowpea, mungbean, azuki bean, bambara groundnut, moth bean, and rice bean. Cultivated primarily in warm temperate and tropical regions worldwide, these crops are renowned for their grains rich in easily digestible proteins. Moreover, they serve diverse agricultural purposes, including as forage, green manure, and cover crops. The crops' short life cycle renders them suitable for catch cropping, intercropping, mixed cropping, or relay cropping. Despite the development of improved cultivars, the full yield potential of various *Vigna* crops is hindered by persistent challenges posed by biotic and abiotic stresses[17].

The *Vigna reflexo-pilosa* var. *glabra* (Roxb.) N.Tomooka & Maxted contributes to the study of genomic consequences of polyploidization within the genus *Vigna* due to its distinctive characteristics within the genus, known for its polyploidy nature, setting it apart from the predominantly diploid composition of others. *V. reflexo-pilosa* is an allotetraploid formed through hybridization from two genome donors, exhibiting distinct differences in flower, leaf, seed size, and other characteristics compared to the species in genus *Vigna*. Additionally, it demonstrates strong resistance to several insect pests and diseases, including bruchids, bean fly, powdery mildew, and cucumber mosaic virus[18,19]. While not widely distributed as a food crop, there are cases where research has been conducted to introduce genes of *V. reflexo-pilosa* into mungbean[20]. Also, understanding how allotetraploidization has shaped the genetic diversity and adaptive capacity of *Vigna* species has important implications for crop improvement. The diversity within this genus provides an opportunity to study the genetic outcomes associated with allotetraploidization. Some species within the genus, such as *Vigna hirtella* Ridl. and *Vigna trinervia* (B.Heyne ex Wight & Arn.) Tateishi & Maxted, have been proposed as potential ancestors of *V. reflexo-pilosa*[21,22], highlighting the importance of *Vigna* for studying these evolutionary dynamics. A phylogenetic analysis based on simple sequence repeats (SSRs) showed that specific taxa of *V. hirtella* and *V. trinervia* contributed their genomic components to *V. reflexo-pilosa*[22]. Furthermore, morphological variation, with *V. reflexo-pilosa* being nearly twice the size of *V. hirtella* and *V. trinervia*, suggests that polyploidization may play a role[23].

We aimed to identify the potential genome donors to *V. reflexo-pilosa*. Using chloroplast genome sequence from RNA-seq of *Vigna* species and high quality reference nuclear genome sequence were compared to elucidate gene-level evidence for allopolyploidization within the genus *Vigna* Through these efforts, we expect to gain a deeper understanding of the complexities involved in polyploidization and genome evolution.

Results

Consensus sequences of the chloroplast genomes of 23 *Vigna* accessions

The newly constructed phylogenetic tree, utilizing consensus sequences from chloroplasts, exhibited a correspondence with major clades and consistent interorganism relationships[21,24]. This alignment provides robust evidence supporting the accuracy and reliability of RNA-seq data in capturing evolutionary signals (Fig. 1). Approximately 140 genes were predicted in each consensus sequence representing different accessions. Among these, 40 genes were found to be covered by the RNA-seq reads. When gene-level nucleotide diversity (Pi) values were calculated for these common regions, rpl33 showed the highest value of 0.01694 (Fig. 2). In addition, we constructed a phylogenetic tree using the given sequences. The results of comparing this tree to the phylogeny result of consensus sequences revealed a Normalized Robinson-Foulds (nRF) score of 0.20 and a Robinson-Foulds (RF) score of 8.0. The Maximum Robinson-Foulds (maxRF) score was found to be 40.0. Additionally, both the source tree (src-br +) and the reference tree (ref-br +) demonstrated branch support values of 0.90.

V. trinervia, previously identified as the ancestor of *V. reflexo-pilosa*, was confirmed to be very closely related. In addition, *V. hirtella*, newly analyzed in this study, was found to belong to the *Angulares* group[24], a section of Asian *Vigna* that encompasses *Vigna angularis*, *Vigna riukiensis*, *Vigna minima*, *Vigna umbellata*, and *Vigna nepalensis*. However, it showed a clear genetic distance from the other five species.

The chloroplast genomes were assembled with sizes of 153,169 bp for *V. reflexo-pilosa*, 151,161 bp for *V. trinervia*, and 151,564 bp for *V. hirtella* (Fig. 3). Consensus sequences, generated by RNA-seq measured 151,185 bp, 151,151 bp, and 151,211 bp, respectively. Although some variations were observed, such as the presence of psbM in the *V. hirtella* de novo assembly result, which was not confirmed in the consensus sequence, overall the sequence of key components was well matched at the genetic level. Using BLAST alignment against the chloroplast genome de novo assembly results of *V. trinervia*, *V. hirtella*, and *V. reflexo-pilosa*, 38 genes derived from the *V. trinervia* chloroplast consensus sequence, 34 genes from the *V. hirtella* chloroplast consensus sequence, and 37 genes from the *V. reflexo-pilosa* chloroplast consensus sequence showed the highest sequence similarity to their respective genomes among the 23 accessions used for Pi calculations (Supplementary Table S1).

Whole genome de novo assembly and annotation

Both *V. hirtella* and *V. trinervia* were subjected to de novo assembly using Illumina sequencing, incorporating paired-end and mate-pair library preparation methods. The resulting assemblies showed satisfactory contiguity, as indicated by N50 values of 209.5 Kb and 496.6 Kb, respectively. Repeat profiling of the assembled genomes revealed high similarity in their repetitive element profiles, with retroelements accounting for approximately 10 to 11 percent of the total genome sequence (Supplementary Table S2). The gene catalogs of *V. hirtella* and *V. trinervia* were assembled using a combination of ab initio and homology-based methods, complemented

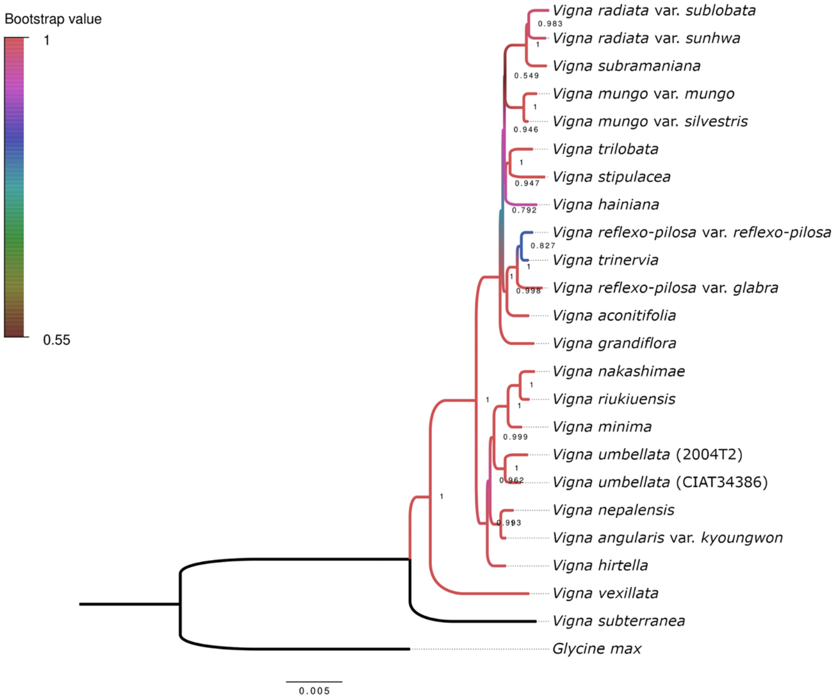

Figure 1. Phylogenetic reconstruction based on single copy chloroplast genes, using protein inference derived from transcriptomic data. This tree elucidates the evolutionary relationships among the 23 *Vigna* accessions and provides insights into their possible genetic neighborhoods.

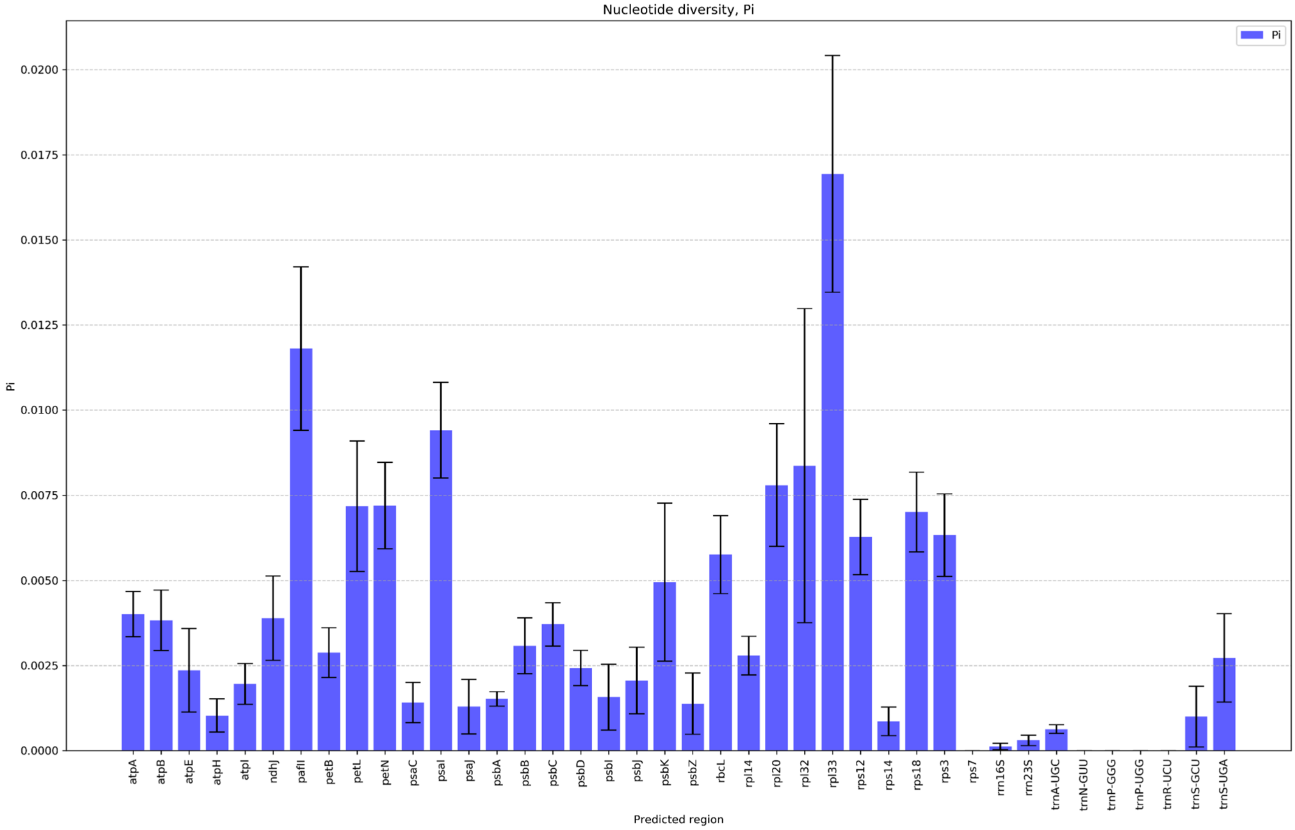

Figure 2. Nucleotide diversity (Pi) values with standard deviation for 40 predicted genes among 23 *Vigna* accessions. The diversity level among species for rpl33 appears to be the highest, while rps7, trnN-GUU, trnP-GGG, trnP-UGG, and trnR-UCU have been identified as well conserved within the genus *Vigna*.

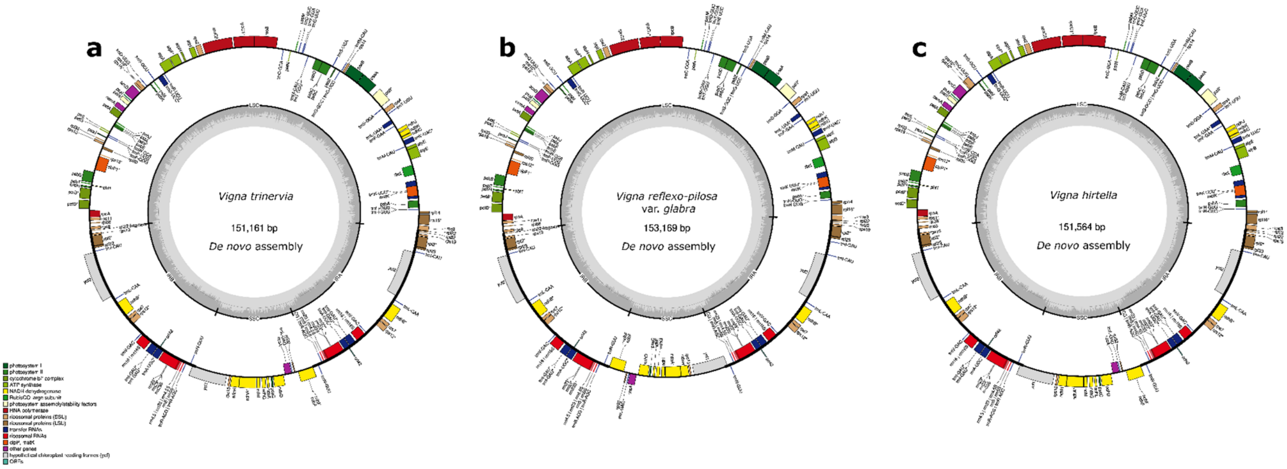

Figure 3. De novo assembly and gene prediction results of the chloroplast genomes of (**a**) *V. trinervia*, (**b**) *V. reflexo-pilosa,* and (**c**) *V. hirtella*. Genes on the inner circle are transcribed in a clockwise direction, while those on the outer circle are transcribed counterclockwise. The darker gray shade on the inner circle indicates the presence of GC content, while the lighter gray shade represents AT content.

by transcriptomic data. A comparison of gene abundance in each species revealed 21,220 genes in *V. hirtella* and 23,546 genes in *V. trinervia*. The observed distribution patterns of mRNA length and coding sequence length (CDS) were comparable for both species, providing a robust basis for subsequent investigations (Fig. 4a). Benchmarking of Universal Single-Copy Orthologs (BUSCO) revealed that more than 90% of the genes were complete in both species (Fig. 4b).

Comparative analysis of *V. hirtella*, *V. trinervia*, and *V. reflexo-pilosa*

Gene family evolution of the three Vigna species

To elucidate the allopolyploid evolution of *Vigna* species, we performed comparative genomic analyses between our sequenced assemblies of *V. trinervia*, *V. hirtella* and the previously assembled *V. reflexo-pilosa* from our previous research[21]. Using the eggNOG database, we annotated the predicted gene catalog encompassing the three *Vigna* species and assigned corresponding eggNOG IDs indicating the respective gene families. Comparative analysis of copy numbers within each gene family revealed that *V. reflexo-pilosa* had a higher copy number distribution compared to the other *Vigna* species (Fig. 5a). Gene families within *V. reflexo-pilosa* that exhibited a twofold increase in copy number compared to the other *Vigna* species may represent the residual signatures of a polyploidization event. Study of these amplified gene families is essential to unravel the genomic implications of such polyploidization, including exploration of the affected biochemical pathways and elucidation of their functional consequences. Our analysis revealed 1221 gene families that fit this scenario. Further annotation of these gene families using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database revealed specific pathways that were preferentially conserved after polyploidization[25] (Fig. 5b). Notably, the "ribosome" and "spliceosome" pathways had twice as many copies in *V. reflexo-pilosa*. The observed trend of amplified ribosomal DNA (rDNA) is consistent with previous research[26].

Comparative synteny analysis among three Vigna species and species tree construction

To validate the previously proposed donor species of *V. hirtella* and *V. trinervia*, we split the *V. reflexo-pilosa* genome into constituent A and B genomes (Fig. 6a). Self-synteny analysis of the *V. reflexo-pilosa* genome revealed synteny blocks that may indicate a past polyploidization event, with a modal synonymous substitution rate (Ks) of 0.064. Interestingly, synteny analysis between *V. hirtella* and *V. trinervia* revealed a comparable modal Ks value of 0.05, which is close to the self-synteny Ks value obtained for *V. reflexo-pilosa*. Comparative analysis of genetic divergence between *V. trinervia* and *V. reflexo-pilosa*, and between *V. hirtella* and *V. reflexo-pilosa*, yielded initial peak values of 0.015 and 0.005, respectively (Fig. 6b). These results suggest a possible genetic relationship: *V. trinervia* is closely aligned to the 'A' genome of *V. reflexo-pilosa*, while *V. hirtella* shows a greater affinity to the 'B' genome of *V. reflexo-pilosa* (Supplementary Figure S2). This finding provides compelling evidence to support the hypothesis that *V. hirtella* may be a plausible candidate as a donor species for *V. reflexo-pilosa*. Furthermore, using the orthologous genes identified through the bioinformatics pipeline, we constructed a phylogenetic tree using a Bayesian inference approach with the BEAST software[27], which strongly suggests that *V. hirtella* is the most likely candidate for the donor genome, showing a closer genetic proximity to the B genome of *V. reflexo-pilosa* compared to other species within the genus *Vigna* (Fig. 6c).

When *V. hirtella* and *V. trinervia* sequences were mapped to the *V. reflexo-pilosa* genome, it was observed that despite average coverage depths of 39.4× and 50.2×, respectively, the contig-wise peak frequency ranges were 0–16 and 46–75 for *V. hirtella* and 0–16 and 55–106 for *V. trinervia* (Fig. 7a). Furthermore, it was found that *V. hirtella* and *V. trinervia* sequences aligned to each contig in a complementary manner (Fig. 7b, Supplementary Figure S2, and Supplementary Figure S3), and the total size of the distinct *V. hirtella*-dominant and *V. trinervia*-dominant

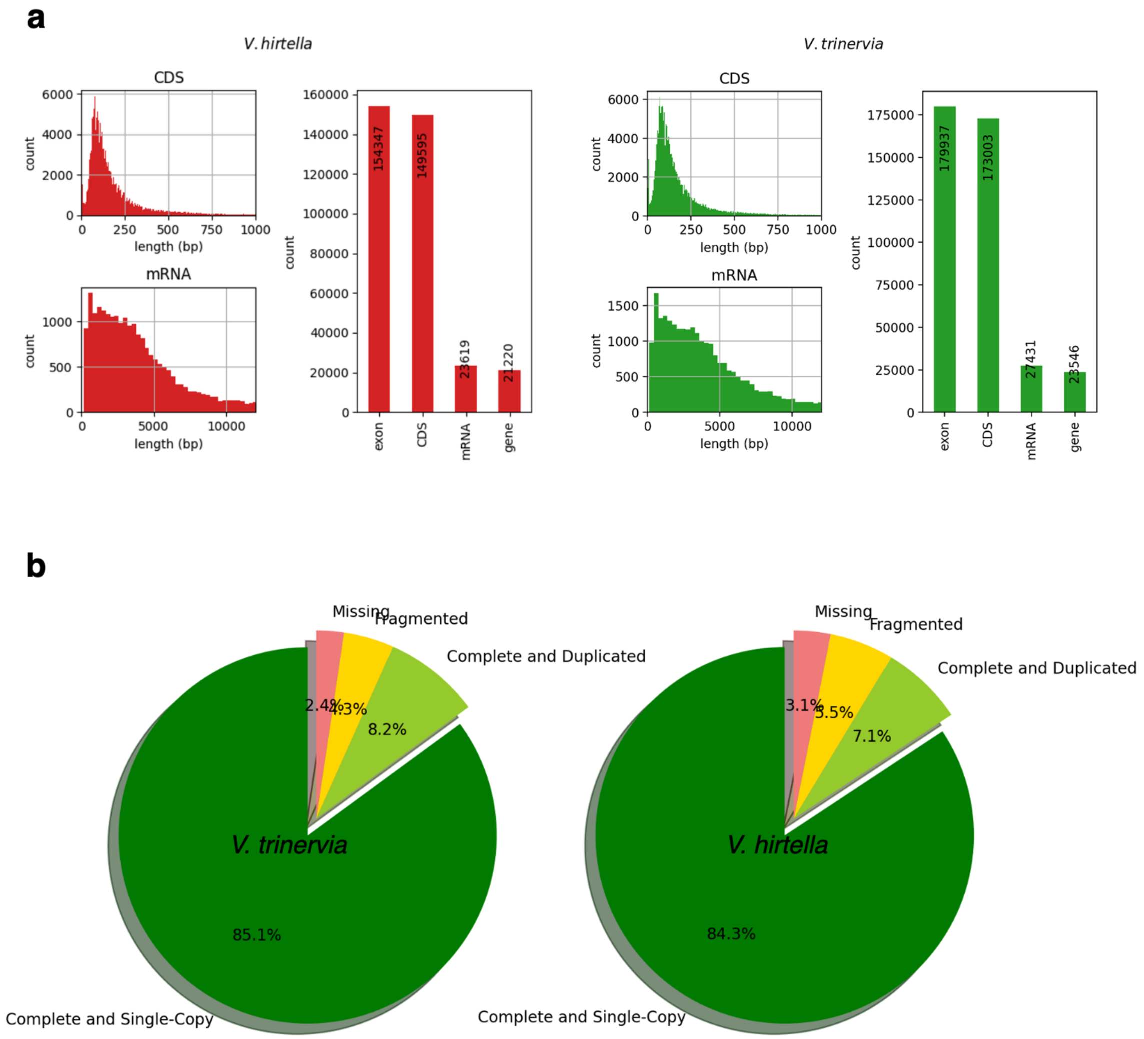

Figure 4. Comparative analysis of the gene prediction profiles of *V. hirtella*. and *V. trinervia* (**a**) A comprehensive visualization of the predicted gene catalogs of the *V. hirtella* and *V. trinervia* genomes, allowing a detailed exploration of their genomic landscapes. (**b**) Evaluations of the quality and completeness of the genome assemblies, using the Benchmarking Universal Single-Copy Orthologs (BUSCO) framework as a reliable metric for assessment.

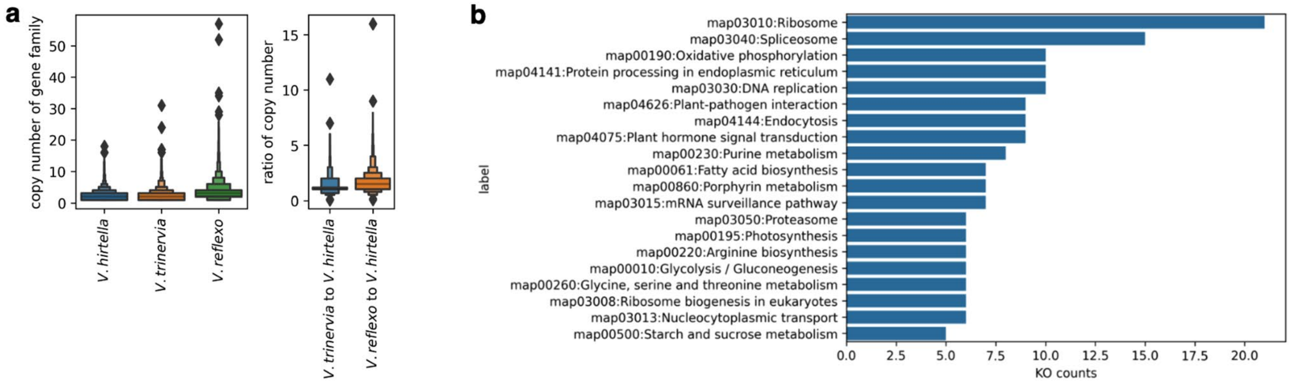

Figure 5. Copy number amplification in *V. reflexo-pilosa*. (**a**) Distribution of gene family copy numbers and copy number ratios among *V. trinervia*, *V. hirtella*, and *V. reflexo-pilosa*. (**b**) KEGG pathway representation of the gene family showing a twofold increase in *V. reflexo-pilosa* compared to *V. hirtella* and *V. trinervia*.

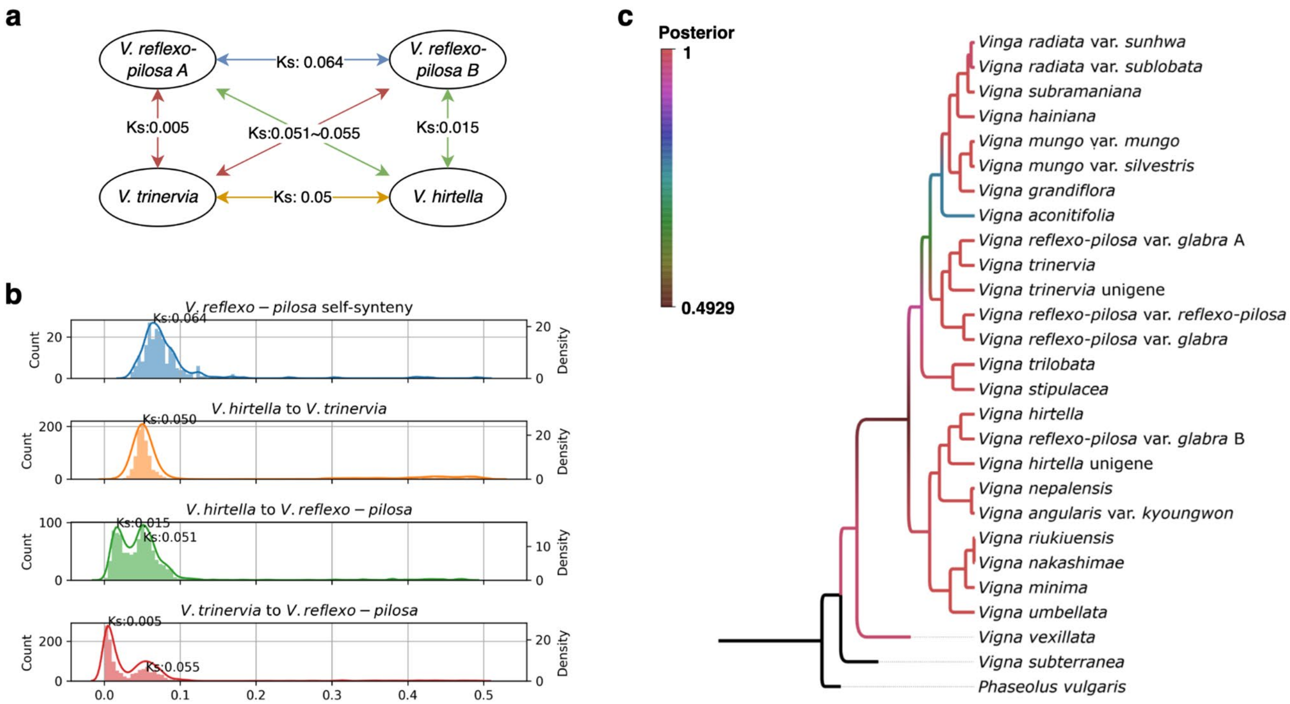

Figure 6. Demonstration of polyploidization of *V. reflexo-pilosa* derived from *V. hirtella* and *V. trinervia*. (**a**) A diagram showing the process of polyploidization of *V. reflexo-pilosa*, supported by Ks values. (**b**) Distribution of Ks values showing self-genome synteny in *V. reflexo-pilosa*, as well as synteny between *V. hirtella* and *V. reflexo-pilosa*, and *V. trinervia* and *V. reflexo-pilosa*. (**c**) A Bayesian species tree derived from genome and transcriptome data of different *Vigna* species, with the genome of *V. reflexo-pilosa* divided into A and B sub-genomes based on self-synteny analysis.

contigs was approximately 309.5 Mb and 332.0 Mb, respectively (Table 1). Based on these results, it was possible to determine the ancestor from which the assembly-generated contigs of *V. reflexo-pilosa* is originated and contigs derived from *V. hirtella* appeared to be less capable of forming long connections compared to those from *V. trinervia*.

Discussion

A phylogeny constructed from consensus sequences of chloroplast genomes is congruent with the established species phylogeny, despite the limitations of relying on the sequence used as a reference, which may not fully account for genome-wide structural variants. However, the uniparental inheritance of chloroplasts, predominantly from the maternal lineage, is a limitation in estimating all ancestors involved in hybridization or polyploidization events. In this study, we could only confirm one of the genome donors of *V. reflexo-pilosa* as *V. trinervia* or its closely related species. In our previous study[21], it was proposed that the B genome donor of *V. reflexo-pilosa* belongs to the *Angulares* section of Asian *Vigna*; however, it is expected to exhibit a distant diversity point compared to other *Angulares* species. Given that, in our phylogenetic study using the consensus sequences of chloroplast genome, *V. hirtella* fulfills the criteria, it can be considered as another plausible candidate for the genome donor of *V. reflexo-pilosa*. Since both *V. hirtella* and *V. trinervia* are found in the Southeast Asian region, natural hybridization between these species, with *V. trinervia* as the maternal ancestor and *V. hirtella* as the paternal donor, seems plausible.

The de novo genome assembly and annotation of *V. hirtella* and *V. trinervia* using Illumina sequencing, along with RepeatMasker and our in-house analysis pipeline, provided insights into the genetic composition of these species. The similarity in repeat profiles, gene counts, mRNA length, and coding sequence length confirmed the reliability of the sequencing and assembly processes.

The observed higher gene family copy number in *V. reflexo-pilosa* suggests the presence of residual signatures from a previous polyploidization event. The enrichment of specific pathways in *V. reflexo-pilosa*, such as "ribosome" and "spliceosome" processes, provides insight into the impact of polyploidization on genome functionalization. The increased rDNA copy number in polyploid *Vigna* species may confer functional advantages that allow them to maintain higher levels of ribosome biosynthesis under stress conditions, potentially increasing their resilience to adverse environments[28].

The synteny analysis, together with Ks value calculations, provides evidence that *V. hirtella* is a close relative and likely donor species to *V. reflexo-pilosa*. This finding is consistent with previous studies using genotype patterns derived from simple sequence repeat (SSR) markers[22]. The phylogenetic tree based on Bayesian inference further supports this hypothesis. However, we acknowledge that inferences based on genetic correlation can be affected by factors such as selective pressure, mutation rates, and genetic drift[29]. Therefore, further investigation

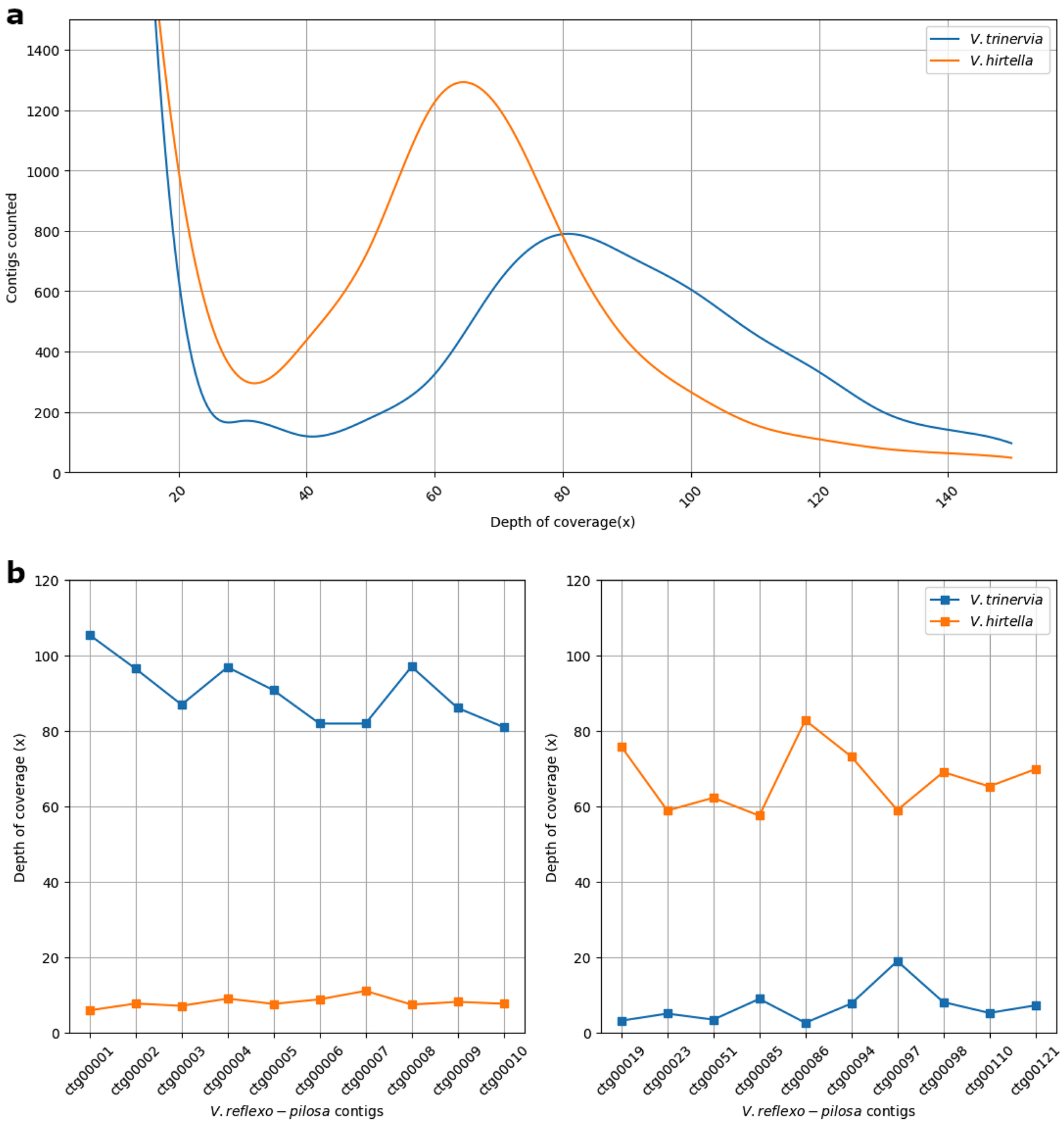

Figure 7. Paired-end reads from *V. hirtella* and *V. trinervia* were mapped to the *V. reflexo-pilos* genome. (**a**) The number of contigs observed varies with coverage, with a common trend of increased contig number occurring at depths below ×16. (**b**) Top10 contigs, ranked by length where *V. hirtella* and *V. trinervia* appear to be dominant at coverage depth.

Mapped condition over contigs	Contig number	Total size	Average depth (standard deviation) of *V. trinervia*	Average depth (standard deviation) of *V. hirtella*
DOC under 16 × for *V. trinervia* reads	5284	309,549,858		70.4 (40.8)
DOC under 16 × for *V. hirtella* reads	4140	331,975,605	91.8 (35.7)	
DOC under 16 × for both species' reads	11	300,539		

Table 1. Depth of coverage (DOC) was calculated for contigs corresponding to the N90 value in the *V. reflexo-pilosa* assembly results using paired-end data from *V. hirtella* and *V. trinervia*.

with more accessions of *Vigna* species is warranted to gain a deeper understanding of speciation within this genus.

In general, when mapping self-sequences or sequences from related species to the target genome, variations in depth coverage can occur from contig to contig due to factors such as sequencing bias. However, in most cases, although there may be a bias toward lower depths, it tends to follow a Gaussian distribution centered around the average depth of the genome. When ancestral sequences were aligned directly to allopolyploidy, it was observed that the mapping pattern at the contig level was concentrated at much lower depths or slightly higher values than the average depth. These results can be attributed to the fact that sequences originating from each progenitor did not align well with sequences from other progenitors. This suggests that at the genome level of *V. reflexo-pilosa*, the sequences derived from each progenitor are well conserved and maintained in their respective forms. To validate this assumption, a comparison was made using the results of self-synteny analysis in 226 paired regions. The result of this analysis showed that, except for a single contig (Vreflexopilosa_ctg411) where a synteny block was detected within the contig itself, all other regions showed a bias towards higher depths in either *V. trinervia* or *V. hirtella*, supporting the notion that sequences from each ancestor tend to maintain their distinct characteristics within *V. reflexo-pilosa*.

Despite the similarities in genome size and gene prediction results between *V. hirtella* and *V. trinervia* during the assembly analysis using the same method, it was observed that fundamental assembly statistics such as the number of contigs and N50 values, exhibited less favorable results in *V. hirtella* than in *V. trinervia*. When examining the mapping of reads from each progenitor to *V. reflexo-pilosa*'s contigs, it was noted that longer contigs tended to have a higher depth of alignment with *V. trinervia* sequences. Connecting these observations, it implies that there might be specific factor hindering *V. hirtella* from making a substantial contribution to the assembly of longer contigs when utilizing short reads.

Future research should focus on understanding the influence of conserved and expanded gene families and the potential adaptive advantages conferred by increased rDNA copy numbers on the functional dynamics and adaptability of *V. reflexo-pilosa*. The observed rapid and dynamic evolution of the rDNA gene family, similar to previous studies in yeast[30], may play an important role in enhancing the adaptability and domestication processes of plant species[31].

In summary, our research provides genomic evidence for polyploidization within the genus *Vigna* and identifies potential donor species for allotetraploid species through de novo genome assembly. These findings provide valuable insights into the gene clusters affected by polyploidization, which may have important implications for plant adaptability and domestication processes. Thus, our research significantly advances the current understanding of plant evolution and the underlying mechanisms of plant adaptation.

Methods

Plant materials

V. hirtella was newly included in 22 accessions from 18 different *Vigna* species, including both Asian and African domesticated varieties that were mentioned in our previous study[21]. These accessions were collected from various national and international genebanks, namely Chai Nat Field Crops Research Center in Thailand, National Agrobiodiversity Center in Korea, National Institute of Agrobiological Sciences in Japan, National Botanic Garden of Belgium, Australian Collections of Plant Genetic Resources, International Center for Tropical Agriculture in Colombia, International Livestock Research Institute in Kenya, and International Institute of Tropical Agriculture in Nigeria.

Consensus sequence of the chloroplast genome

RNA-seq data for each *Vigna* species sample were aligned to the chloroplast sequence (NC_013843.1) using BWA mem 0.7.17-r1188[32]. Duplicate reads were removed using sambamba v.0.6.8[33] and variant calling was performed using SAMtools 1.9. Variants with a phred score of 30 or higher were used to generate a consensus sequence for *V. radiata* chloroplast DNA using bcftools 1.9. The consensus sequences of each accession were aligned using MAFFT v7.453[34], and a neighbor-joining method with 1000 bootstrap replications was used for phylogenetic analysis on the resulting alignment.

We also performed gene prediction and annotation on the *Vigna* chloroplast consensus sequence using GeSeq 2.03[35]. Subsequently, we filtered regions representing comprehensive coverage by RNA-seq data and calculated nucleotide diversity (Pi) using DnaSP v6.12.03[36]. ETE3 v3.1.3[37] was used to compare phylogenetic trees from consensus sequences and filtered regions.

Based on the constructed phylogenetic tree, we selected specific species and performed de novo assembly of chloroplast genomes using DNA sequences through GetOrganelle 1.7.7.9[38] with default option.

De novo assembly and annotation

To perform de novo assembly, we first estimated the genome size of the sample using Jellyfish v1.1.11[39] with k-mer analysis at 17, 21, and 25 (Supplementary Figure S4). Platanus-allee v2.2.0[40] software was used to perform de novo assembly with paired-end reads and mate pair reads of different insert sizes (350 bp for paired-end reads, 5 kb, and 10 kb for mate pair reads). Scaffolding and gap filling were performed on mate pair reads using SSPACE v2.1.1[41], and the best scaffold was selected based on number of scaffolds, scaffold sum, and N50. A length cutoff was applied to remove short scaffolds (Supplementary Figure S5). The assembly results of the *V. reflexo-pilosa* genome were examined by directly aligning the paired-end reads of *V. hirtella* and *V. trinervia* to assess the overall mapping pattern.

Repeat masking was performed using the RepeatModeler 2.0.4[42] and RepeatMasker 4.1.5[43] pipelines to identify and mask repetitive elements in the assembled genome. RepeatModeler was used to generate a de novo repeat library, which was then used by RepeatMasker to mask repetitive elements in the genome sequence.

For genome-guided transcriptome assembly, RNA reads were mapped to the assembled DNA sequence using Tophat v2.0.13 software[44], and the assembled transcriptome sequence was obtained from the resulting BAM file using Trinity r20140717[45]. Annotation of the assembled DNA sequence and transcriptome sequence data was performed using the Seqping v0.1.33 pipeline[46], which included gene prediction models built using GlimmerHMM v3.0.4, AUGUSTUS v3.2.2, and SNAP 20120517 software[47,48]. Prediction results were combined with the

MAKER v3.01.03[49] annotation program included in the Seqping pipeline. Additional annotation was performed by searching consensus sequences against several databases, including UniProt[50], GO[51,52], InterPro[53], Pfam[54], TIGRFAM[55], and eggNOG[56] using blast v2.6.0 + software[57].

Comparative genomics

A multi-step process was used to identify true orthologs (Supplementary Figure S6). Synteny analysis using MCScanX[58] was performed on the *V. reflexo-pilosa*, *V. trinervia*, and *V. hirtella* genomes to explore syntenic relationships within the reference genomes. Self-synteny analysis was performed on *V. reflexo-pilosa* to detect matching regions, which were then partitioned based on the Ks value. The portion closer to *V. trinervia* was designated the 'A' genome, while the more distant portion was designated the 'B' genome.

Next, BLAST analysis was then performed to assign proteins from the *V. hirtella* and *V. trinervia* genome assemblies to each transcriptome assembly of the 22 *Vigna* accessions from the previous study[21] to identify candidate orthologs. Gene family relationships between the transcriptome assemblies and the assembled genomes were determined using the eggNOG database[56]. Proteins identified as matches in both the BLAST result and the eggNOG database search were classified as true orthologs.

Data availability

Raw sequence reads are deposited in the SRA (BioProject: PRJNA961890). The assembled sequences of *V. reflexo-pilosa, V. trinervia* and *V. hirtella* are available on NCBI with BioSampleID: SAMN34371969, SAMN34371970 and SAMN34371971.

Received: 20 September 2023; Accepted: 13 December 2023
Published online: 22 December 2023

References

1. Anatskaya, O. V. & Vinogradov, A. E. Polyploidy as a fundamental phenomenon in evolution, development, adaptation and diseases. *Int. J. Mol. Sci.* **23**, 3542 (2022).
2. Soltis, D. E., Visger, C. J. & Soltis, P. S. The polyploidy revolution then…and now: Stebbins revisited. *Am. J. Bot.* **101**, 1057–1078 (2014).
3. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005).
4. Madlung, A. Polyploidy and its effect on evolutionary success: Old questions revisited with new tools. *Heredity* **110**, 99–104 (2013).
5. Wolf, D. E., Steets, J. A., Houliston, G. J. & Takebayashi, N. Genome size variation and evolution in allotetraploid *Arabidopsis kamchatica* and its parents, *Arabidopsis lyrata* and *Arabidopsis halleri*. *AoB Plants* **6**, plu025 (2014).
6. *Soybean: Botany, Production and Uses.* (CABI Publishing, 2010).
7. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
8. Doyle, J. J., Doyle, J. L., Rauscher, J. T. & Brown, A. H. D. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (*Glycine*subgenus*Glycine*). *New Phytol.* **161**, 121–132 (2004).
9. Birchler, J. A. & Veitia, R. A. The gene balance hypothesis: From classical genetics to modern genomics. *Plant Cell* **19**, 395–402 (2007).
10. De Storme, N. & Mason, A. Plant speciation through chromosome instability and ploidy change: Cellular mechanisms, molecular factors and evolutionary relevance. *Curr. Plant Biol.* **1**, 10–33 (2014).
11. Freeling, M., Scanlon, M. J. & Fowler, J. E. Fractionation and subfunctionalization following genome duplications: Mechanisms that drive gene content and their consequences. *Curr. Opin. Genet. Dev.* **35**, 110–118 (2015).
12. Dodsworth, S., Chase, M. W. & Leitch, A. R. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms?. *Bot. J. Linn. Soc.* **180**, 1–5 (2015).
13. Parkin, I. A. P. *et al.* Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics* **171**, 765–781 (2005).
14. Gaeta, R. T., Pires, J. C., Iniguez-Luy, F., Leon, E. & Osborn, T. C. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell* **19**, 3403–3417 (2007).
15. Lim, K. Y. *et al.* Rapid chromosome evolution in recently formed polyploids in Tragopogon (Asteraceae). *PLoS One* **3**, e3353 (2008).
16. Li, Q. *et al.* Unbiased subgenome evolution following a recent whole-genome duplication in pear (*Pyrus bretschneideri*Rehd.). *Hortic. Res.* **6**, 34 (2019).
17. *Alien Gene Transfer in Crop Plants, Volume 2.* (Springer New York).
18. Tateishi, Y. Contribution to the genus Vigna (leguminosae) in Taiwan. *Sci. Rep. Tohoku Univ. 4th Ser. (Biology)* **38**, 335–350 (1984).
19. Tomooka, N., Wa, Y. E., Lairungreang, C. & Arasook, C. T. V. Collection of wild ceratotropis species on the nansei. https://www.jircas.go.jp/sites/default/files/publication/jarq/26-3-222-230_0.pdf (1992).
20. Somta, P., Seehalak, W. & Srinives, P. Development, characterization and cross-species amplification of mungbean (*Vigna radiata*) genic microsatellite markers. *Conserv. Genet.* **10**, 1939–1943 (2009).
21. Kang, Y. J. *et al.* Genome sequence of mungbean and insights into evolution within Vigna species. *Nat. Commun.* **5**, 5443 (2014).
22. Chankaew, S. *et al.* Detection of genome donor species of neglected tetraploid crop *Vigna reflexo-pilosa* (créole bean), and genetic structure of diploid species based on newly developed EST-SSR markers from azuki bean (*Vigna angularis*). *PLoS ONE* **9**, e104990 (2014).
23. Heslop-Harrison, J. S. P., Schwarzacher, T. & Liu, Q. Polyploidy: Its consequences and enabling role in plant diversification and evolution. *Ann. Bot.* **131**, 1–10 (2023).
24. Tomooka, N. *The Asian Vigna: Genus Vigna Subgenus Ceratotropis Genetic Resources* (Springer, 2002).
25. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
26. Rosato, M., Moreno-Saiz, J. C., Galián, J. A. & Rosselló, J. A. Evolutionary site-number changes of ribosomal DNA loci during speciation: Complex scenarios of ancestral and more recent polyploid events. *AoB Plants* **7**, 135 (2015).
27. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
28. Grummt, I. The nucleolus—Guardian of cellular homeostasis and genome integrity. *Chromosoma* **122**, 487–497 (2013).
29. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
30. Sultanov, D. & Hochwagen, A. Varying strength of selection contributes to the intragenomic diversity of rRNA genes. *Nat. Commun.* **13**, 7245 (2022).
31. Gepts, P. The contribution of genetic and genomic approaches to plant domestication studies. *Curr. Opin. Plant Biol.* **18**, 51–59 (2014).

32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
34. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
35. Tillich, M. *et al.* GeSeq—versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
36. Rozas, J. *et al.* DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017).
37. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
38. Jin, J.-J. *et al.* GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
39. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
40. Kajitani, R. *et al.* Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nat. Commun.* **10**, 1702 (2019).
41. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
42. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
43. Smit, A. F. A., Hubley, R. & Green, P. 2013–2015. RepeatMasker Open-4.0. (2021).
44. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
45. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
46. Chan, K.-L. *et al.* Seqping: Gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data. *BMC Bioinform.* **18**, 1426 (2017).
47. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
48. Stanke, M. *et al.* AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
49. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
50. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
51. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
52. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, (2023).
53. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
54. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
55. Haft, D. H. *et al.* TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29**, 41–43 (2001).
56. Huerta-Cepas, J. *et al.* eggNOG 50: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
57. Camacho, C. *et al.* BLAST: Architecture and applications. *BMC Inform.* **10**, 421. Preprint at (2009).
58. Wang, Y. *et al.* MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).

Acknowledgements

This work was carried out with the support of the "Cooperative Research Program for Agriculture Science and Technology Development" (Project No. RS-2021-RD009467).

Author contributions

J.L. designed the experiment and pipelines of bioinformatics; Y.J.K, H.P., S.S., and T.L. performed the bioinformatic analysis; J.H. and M.Y.K. led the study of *Vigna* speciation; S.-H.L. initiated and coordinated the research project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-49908-2.

Correspondence and requests for materials should be addressed to S.-H.L.