SI Appendix

Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis and allergens

Xiaoping Chen[1,$], Hongjie Li[2,$], Manish K. Pandey[3,$], Qingli Yang[4,8,$], Xiyin Wang[5,$], Vanika Garg[3], Haifen Li[1], Xiaoyuan Chi[4], Dadakhalandar Doddamani[3], Yanbin Hong[1], Hari D. Upadhyaya[3], Hui Guo[5], Aamir W. Khan[3], Fanghe Zhu[1], Xiaoyan Zhang[2], Lijuan Pan[4], Gary J. Pierce[5], Guiyuan Zhou[1], Katta AVS Krishnamohan[3], Mingna Chen[4], Ni Zhong[1], Gaurav Agarwal[3], Shuanzhu Li[2], Annapurna Chitikineni[3], Guoqiang Zhang[7], Shivali Sharma[3], Na Chen[4], Haiyan Liu[1], Pasupuleti Janila[3], Shaoxiong Li[1], Min Wang[2], Tong Wang[4], Jie Sun[4], Xingyu Li[1], Chunyan Li[2], Mian Wang[4], Lina Yu[4], Shijie Wen[1], Sube Singh[3], Zhen Yang[4], Jinming Zhao[2], Chushu Zhang[4], Yue Yu[6], Jie Bi[4], Xiaojun Zhang[8], Zhongjian Liu[7,*], Andrew H. Paterson[5,*], Shuping Wang[2,*], Xuanqiang Liang[1,*], Rajeev K. Varshney[3,9,*], Shanlin Yu[4,*]

[1]Crops Research Institute, Guangdong Academy of Agricultural Sciences (GAAS), South China Peanut Sub-center of National Center of Oilseed Crops Improvement, Guangdong Key Laboratory for Crops Genetic Improvement, Guangzhou, China

[2]Shandong Shofine Seed Company, Jiaxiang, China

[3]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

[4]Shandong Peanut Research Institute, Shandong Academy of Agricultural Sciences, Qingdao, China

[5]Plant Genome Mapping Laboratory, University of Georgia, Athens, USA

[6]Macrogen Millennium Genomics Company, Shenzhen, China

[7]Shenzhen Key Laboratory for Orchid Conservation and Utilization, National Orchid Conservation Center of China and Orchid Conservation and Research Center of Shenzhen, Shenzhen, China

[8]College of Food Science and Engineering of Qingdao Agricultural University, Qingdao, China

[9]The University of Western Australia, Crawley, Australia

$ These authors contributed equally to this work.

*Correspondence should be addressed to S.Y. (yshanlin1956@163.com), R.V. (r.k.varshney@cgiar.org), X.L. (liang-804@163.com), S.W. (wsp@shofine.com), A.P. (paterson@uga.edu) or Z.L. (liuzj@sinicaorchid.org).

SI Text

1. Sequencing and assembly of *Arachis duranensis*

1.1 Plant material

Arachis duranensis (AA 2n=2x=20) is the progenitor species of the cultivated peanut[1,2] (**Fig. S1**). The *A. duranensis* (represented as accession PI475845) was sequenced by Illumina HiSeq2500 sequencing platform. Genomic DNA was extracted from the etiolated leaves of 20-day-old plants growing in dark chamber using the CTAB method[3].

1.2 Illumina shotgun sequencing

Genomic DNA was isolated from caulicle, leaf and root by standard molecular biology techniques. Subsequently, short-insert libraries (250-bp, 500-bp & 800-bp) and long-insert libraries (2-kb, 5-kb, 10-kb & 20-kb for BP) were constructed using the standard protocol provided by Illumina (San Diego, USA). Paired-end sequencing with whole genome shotgun sequencing strategy was performed using the Illumina HiSeq 2500 platform. We finally obtained ca. 229.94G reads for next filter step (**Table S1)**.

1.3 *De novo* assembly of the *A. duranensis* genome

The schematic strategy for *de novo* assembly is displayed in **Fig. S2**. Sequencing errors will largely disturb the short-read assembly algorithms. We therefore utilized several highly stringent filtering steps to remove low-quality reads as follows: (1) reads of short-insert libraries were trimmed of 4 low-quality bases at both ends, and reads of long-insert libraries were trimmed of 3 low-quality bases; (2) for long-insert libraries, duplicated reads were filtered out; (3) we also examined individual reads in all lanes, and discarded reads with 10 or more Ns (no sequenced bases) and low-quality bases.

We finally obtained 159.07G filtered reads for genome assembling. We employed SOAPdenovo2[4] (version 2.04.4) with optimized parameters (pregraph -K 79 -p 16 -d 5; scaff -F -b 1.5) to construct contigs and original scaffolds. This paired-end information was subsequently applied to link contigs into scaffolds in a stepwise manner. Several intra-scaffold gaps were filled by local assembly using the reads in a read-pair where one end uniquely mapped to a contig whereas the other end was located within a gap. Subsequently, SSPACE[5] (version 2.0; using core parameters "-k 6 -T 4 -g 2") was used to link the SOAPdenovo2 scaffolds. Overall, various assembly software were employed to generate a draft genome of *A. duranensis* consisting of 8,173 scaffolds with a total of 1,051,523,805 bp (avg. size: 128,658; N50 size: 649,840) and 90,568 contigs (N50 size: 29,584) (**Table 1** and **Table S2**). Out of 8,173 scaffolds, 3,996 with length ≥2 Kb account for 1.048 Gb of the genome (**Table S3**).

1.4 Evaluation of the assembly

1.4.1 PCR amplification

We evaluated the *A. duranensis* assembled genome using PCR method. A total of 411 genomic fragments from the assembled genome were randomly selected for designing PCR primers. Of the 411 pairs of primers, ~89% can be amplified the right size of product from the genomic DNA of PI475845 (**Table S4** and **Fig. S3**). All primers used in this study were provided in **Dataset S1**.

1.4.2 Per-base accuracy of read data and sequence depth

The accuracy of a genome assembly depends partially on the high quality of sequenced reads, which has a great impact on subsequent analyses. Base errors in the sequenced fragments can not only lead to the deviation of assembly, but also result in the incorrect annotation of functional elements in downstream analyses. The read length and quality distributions were thus explored **(Fig. S4).** Nearly all reads below

1000 bp have high quality (Q>20). The high-quality data guarantees the single base accuracy of the assembled genome and the correct annotation of functional elements like protein-coding genes, transcription factors and small RNAs. The sequencing depth of 93.27% genomic regions was ≥ 10x and its peak locates at 48x (**Fig. S5**), indicating that these regions had high single-base accuracy[6].

1.4.3 EST and Transcriptome Sequence Assembly (TSA) mapping

The gene coverage of the assembled genome was comprehensively evaluated using available transcript sequence tags or ESTs. We used the RNA-seq data[7] generated in-house and downloaded from the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/Traces/sra/). We aligned the transcripts to the genome using SSAHA2[8] with default parameters except for '-best 1'. A total of 50,281 (approximately 99% of the predicted genes) genes were supported by at least one transcripts (**Fig. S6**).

1.5 Estimation of the genome size based on 25-mer analysis

The genome size was estimated based on the K-mer distribution using ~79 Gb of high-quality short reads. A k-mer refers to a total number of sub-sequences of length k which could be obtained from a sequenced DNA read. The genome size was evaluated using the total length of sequence reads divided by sequencing depth. To estimate the sequencing depth, the frequency of each 25-mer were calculated from the whole genome sequenced reads. We used the algorithm: $(N \times (L - K + 1) - B)/D = G$, where N is the total sequence read number, L is the average length of sequence reads and K is K-mer length, defined as 25 bp here, B is the total number of low frequency 25-mer, G denotes the genome size, and D is the overall depth estimated from K-mer distribution (**Table S7**). An average of 57.14x read depth was obtained with an estimated genome size of 1,381,794,909 bp, consistent with the prior data[9].

2. Genome annotation

2.1 Gene prediction

To annotate the *A. duranensis* genome, we used an automated genome annotation pipeline MAKER[10] which aligns and filters EST and protein homology evidence, produces *de novo* gene prediction, infers 5' and 3' UTR, and integrates these data to generate final downstream gene models with quality control statistics. Several iterative runs of MAKER were used to produce the final gene set. In total, 50,324 gene models for *A. duranensis* were predicted in this study (**Table 1**).

2.2 Gene function annotation

All predicted protein sequences were functionally annotated using the BLAST+ (version 2.2.27) with a threshold E-value of 1e-5 against a variety of protein and nucleotide databases, including the NCBI nucleotide (NT), the non-redundant protein (NR), the Conserved Domain Database (CDD)[11], the UniProtKB (www.uniprot.org), Pfam[12,13] and the Gene Ontology (GO)[14]. The *A. duranensis* genes were also mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps of KEGG databases[15]. To infer functions for the predicted genes, InterProScan[16,17] was used to search the predicted genes against the protein signature from InterPro with default parameters. Fifteen gene sets from legumes, oilseed crops and other plant species were used for comparative analysis (**Table S8**). A Cytoscape plugin BiNGO was used for enrichment analysis with hypergeometric test and Benjamini multiple testing correction at a significance level of 0.01[18].

2.3 Identification of gene and transcription factor families

Comparative analysis of gene family evolution including expansion, contraction, formation or extinction can reveal evolutionary events underlying species adaptation[19]. The software OrthoMCL (version 2.09)[20] was employed to identify orthologous gene families in the *A. duranensis* genome. To cluster protein-coding genes into gene families, pairwise sequence similarity analysis was performed using BLASTP with an E-value cutoff of 1e-5 and a minimum aligned coverage of 50%. The reciprocal best hit matrix served as the basis for ortholog definition using OrthoMCL. The gene sets used in this study are listed in **Table S8**. A total of 832,953 sequences from sixteen plant species were grouped into 54,384 gene clusters, of which 4,575 clusters contained 237,686 genes common to all sixteen genomes, and 1,423 were specific for *A. duranensis*, suggesting that new gene families may have emerged after *Arachis* divergence from other legumes ~50 Mya[21]. These specific clusters are comprised of 16,472 genes, more than in other examined species except canola (**Table S13** and **Fig. S19**). Gene Ontology (GO) annotation indicated differentially enriched functional categories in peanut-specific families (**Fig. S20** and **S21**), suggesting that new gene families may reflect *Arachis* speciation and adaptation to specific habitats, for example by geocarpy. Legumes shared 6,508 (114,289 genes) families (**Fig. S15**), while 8,347 (130,529 genes) and 7,117 (113,667 genes) families were shared with oilseeds and other distantly related species, respectively (**Fig. S16** and **S17**). Shared and unique gene families are shown in **Figs. S15-S17**.

The gene numbers of orthologous families were used to determine the family size by counting the incorporated *A. duranensis* genes for each cluster. We compared the *A. duranensis* gene family size relative to corresponding gene family size in other plant species examined. The number difference of the gene family size and gene copy number were calculated. Then, the median of the *A. duranensis* gene count was determined and a polynomial fit of these values was computed using locally-weighted polynomial regress using an R stats package (http://stat.ethz.ch/R-manual/R-patched/library). A comparison of *A. duranensis* gene family size relative to corresponding gene family size in soybean and *Medicago* was presented in **Table S14**, indicating that approximately 56% of families showed no change in size between *A. duranensis* and soybean, while 73% between *A. duranensis* and *Medicago*, suggesting that expansion and contraction of *A. duranensis* gene families are different from other legumes.

Transcription factors (TFs) can regulate the expression of genes at the transcriptional level. For the identification of known TFs in *A. duranensis*, TFs from other species were retrieved from PlantTFDB (http://planttfdb.cbi.pku.edu.cn/) (**Dataset S5**). For *A. duranensis*, we utilized the predicted gene set against the PlantTFDB databases using BLASTP with an E-value cutoff of 1e-5. A total of 5,251 TFs were identified in *A. duranensis*, consisting of 58 families, representing 10.43% of predicted protein-coding genes (**Dataset S5**). Particularly enriched are TF families such as B3, bHLH, C2H2, C3H, ERF, G2-like, HD-Zip, M-type, YB-related, TCP, Trihelix and WRKY.

2.4 Identification of non-coding RNAs

Non-coding RNAs include highly abundant and functionally important RNAs. In this study non-coding RNA genes refer to four different types: transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA (miRNA) as well as small nuclear RNAs (snRNAs). Non-coding RNAs were annotated by aligned our assembly to against the Rfam database (version 11.0)[22]. Three RNA prediction programs including tRNAscan-SE, RNAmmer and INFERNAL were used to predict the non-coding RNAs in *A. duranensis*. The tRNAs were predicted using the tRNAscan-SE[23], rRNAs were identified using the RNAmmer[24], snRNAs were annotated using the INFERNAL (version 1.0)[25] and other non-coding RNA genes were annotated by aligning the genome sequences against Rfam database (version 11.0). Conserved miRNAs were identified by mapping all entries in miRBase against the assembled genome. Novel miRNAs were identified using miREAP[26].

In *A. duranensis*, we predicted a total of 913 tRNAs with an average length of ~73 bp; 115 rRNAs with an average length of ~1 kb, including 5S (61), 5.8S (17), 18S (21) and 28S (16) as well as 202 snRNAs with an average length of ~127 bp **(Table**

S15). A total of 816 miRNAs, including 801 conserved belonging to 96 families (**Tables S15-S16; Dataset S6**) i.e., more than soybean (390 genes, 85 families) *Medicago* (512 genes, 101 families) (miRBase release 21).

2.5 Annotation of repetitive sequences and transposon elements

We examined the genomic positions of the repeats that were classified as Long Terminal Repeats (LTR), Long Interspersed Nuclear Elements (LINE), Short Interspersed Nuclear Elements (SINE) and DNA transposons. Repetitive sequences in *A. duranensis* were identified using the RepeatMasker, Tandem Repeats Finder (TRF)[27] and RepeatModeler open-1.0[28] for homolog and *de novo* prediction, respectively. We screened the genome using RepeatMasker against the RepBase (version 20110920)[29]. The TE sequences were classified according to the unified classification system[30]. Gaps in the sequences were not included when calculating the total TE contents. A total of 20,597 scaffolds were subjected to the TE identification, 90.2% (18,580) of which were identified as TE sequences. The remained scaffolds without TE could be low-copy sequences or contained uncharacterized repeat sequences so far. Approximately 60% of the *A. duranensis* genome were identified to be TE sequences (**Fig. 1** and **Table 2**).

2.6 Dating the insertion time of LTR retrotransposons

LTR retrotransposons are the most common type of TEs in plants and play a vital evolutionary role in the remarkable divergence of genome size in flowering plants[31]. The identity of both ends of LTR can be used to estimate their insertion time in genome[32]. We used CD-HIT program[33] to cluster LTR retrotransposons based on 90 % sequence similarity (-c 0.9). The longest sequence of each cluster was chosen as the representative sequence, and other sequences within the same cluster must cover 90% of the length of the representative sequence (-aL 0.9). Insertion dates were calculated using the Kimura two-parameter method[34] with the mutation rate of 1.3 x 10^{-8}

substitutions per site per year[35]. The insertion times of LTR retrotransposons were dated to observe the activity of these elements in the *A. duranensis* genome expansion regarding the genome structural variations. The histograms, presented in **Fig. S24,** showed one peak of the insertion times of LTR retrotransposons, revealing these LTR retrotransposons have undergone one burst of amplification ~2 Mya, suggesting that the expansion of the *A. duranensis* genome was relatively recent.

3. Molecular marker development

3.1 Simple sequence repeats (SSRs)

Simple Sequence Repeats (SSRs) in *A. duranensis* were identified using MISA, a MIcroSAtellite identification tool[36] (http://pgrc.ipk-gatersleben.de/misa/). SSRs with di-nucleotide motifs were defined with at least 6 repeats and 5 repeats for tri-, tetra-, penta- and hexa-nucleotide motifs. The maximum number of interrupting nucleotides in a compound SSR was set as 100. The statistics of SSRs (di- up to hexa-nucleotide) in the *A. duranensis* genome was shown in **Table S18**. In total, we detected 105,003 SSRs in *A. duranensis* from which 84,464 SSR primers were designed. The di-nucleotide motif was the most abundant type and accounted for 43.45% of all SSRs, followed by tri-nucleotide (30.54%). In di-nucleotide type, AT motif was the most abundant type. In tri-nucleotide type, AAT was dominant (**Dataset S8**).

3.2 Single nucleotide polymorphism (SNPs)

Reads from six re-sequenced genotypes including two A-genome genotypes (ICG 8123 and ICG 8138) and four B-genome genotypes (ICG 8960, ICG 8209, ICG 13160 and ICG 8206) (**Table S19)** were aligned to the reference genome using the Burrows-Wheeler Aligner program (BWA)[37]. About 70% of reads of A genomes (ICG 8123 and ICG 8138) could be mapped to the *A. duranensis* genome with a threshold that five mismatches are allowed, while ~45% of reads of B genomes (ICG 8960, ICG 8209, ICG 13160 and ICG 8206) could be mapped (**Dataset S10**). SAMtools[38] (version 1.1) was used to call SNPs (**Table S20**). We identified 8,617,722-8,653,808 SNPs against A-genome genotypes and 3,684,730-3,884,005 against B-genome genotypes (**Table S20; Dataset S11**). Fewer SNPs were detected in B-genome genotypes due to fewer mapped reads. Structural variations such as insertions, deletions, copy number variations and inversions for the A- , B- and AB genomes were also identified (**Table S21**).

4. Speciation of peanut A and B subgenome

By performing a trio comparison of the synthetic tetraploid ISATGR 184 and its parents, ICG8123 and ICG8206, we studied the divergence between the subgenomes A and B. Parental reads were mapped to the reference genome and identified SNV between the two parental lines. In total, ~43% of reads from two parental lines were mapped to the reference genome. We filtered the SNPs by read coverage (>4x) and likelihood of second most likely genotype < 0.05. A total of 847676 high quality SNVs were identified between the two parental lines, meaning a mutation rates ~4.5 x 10^{-4} mutations at a base site in each line. Then, we mapped reads from ISATGR to the reference genome. In total, 76.04% of reads were successfully mapped. Genotypes are filtered by read coverage (>20x) and likelihood of second most likely genotype < 0.05. We identified 748802 SNV sites between the two parental line and they were genotyped in the tetraploid species.

5. Evolutionary analysis

The phylogenetic tree was constructed using single-copy orthologous genes shared by *A. duranensis* and fifteen other plant species (soybean, *Medicago*, *Lotus*, pigeonpea, chickpea, common bean, canola, cotton, castor, linseed, *Arabidopsis*, apple, poplar, tomato and rice) using the maximum-likelihood algorithm implemented in MEGA[39]. Colinear genes from *Medicago*[40], soybean[41], and grape[42] were used to locate related evolutionary events. We found evidence that peanut was affected by one lineage-specific event after its divergence from the *Medicago*-soybean lineage. Colinear genes within a genome and between different genomes were inferred by using MCScanX[43]. We adopted soybean genes' CDS in colinearity in its genome to search against peanut scaffold sequences to find best matching pairs of regions > 120 bp in length. Soybean genes were preferred over *Medicago* genes as reference to retrieve peanut homologs in that *Medicago* genes seem to accumulate mutations faster[40]. Genes with tandem duplicates in their respective neighboring 100 kb regions in soybean or from large gene families (with more than 30 genes at BLASTP E-value 1e-10) were removed from the present analysis. We inferred synonymous substitution rates between homologous genes by using the Nei-Gojobori approach implemented in PAML[44]. Peanut coding sequences were aligned with their soybean homologs codon by codon, estimating synonymous substitution rates (*Ks*) between peanut and soybean homologs and between two retrieved peanut CDS. Accordingly, *Ks* between homologs within and among three other plants were estimated. The *Ks* distribution of peanut homologs shows a very prominent peak around *Ks* = 0.02-0.04 (**Fig. 2d**), which suggests a peanut-specific polyploidization. Compared to a previously inferred soybean-specific polyploidization at ~13 Mya[41], the peanut-specific event is much more recent, occurring ~5 Mya.

Reads from different genotypes were aligned to the reference genome by BWA[37]. SAMtools[38] (v1.1) were used to call single nucleotide variations (SNV). SNV sites were compared between parental lines and subgenomes in tetraploids to find likely converted sites and other mutated sites as previously described[45]. SNVs are identified between the two parental lines by mapping reads to the reference genome, with 72.0% and 43.1% of reads from ICG 8123 and ICG 8206 mapped respectively. We filtered SNVs by read coverage (>4x) and likelihood of second most likely genotype < 0.05. A total of 847,676 high quality SNVs were identified between the two parental lines. About 76.04% of the reads from ISATGR 184 are mapped to the reference genome. Genotypes are filtered by read coverage (>20x) and likelihood of second most likely genotype < 0.05. A total of 748,802 SNV sites between the two parental lines were genotyped in the tetraploid species with high accuracy. We found that extensive gene conversion has taken place virtually immediately following polyploid formation, i.e. in the ~3 seed to seed generations that have passed following formation of this neopolyploid by human hands.

6. Synteny analysis

Promer package of MUMmer[46] was used to look for Maximal Unique Matches (MUMs) for the amino acid sequences aligned. The whole genome dot plots for these matches were depicted using the Mummerplot and gnuplot 4.4 patch level 2. The protein sequences of the genomes were compared and clustered using Vmatch[47] with a query and subject coverage of 85 % and 70 % respectively with a minimum match length of 100 and an exdrop of 100. Yn00 of PAML package was used for the identification of duplicated genes in the clusters. The matches were then further provided to i-ADHoRe[48] for the identification of syntenic blocks between two genomes. The coordinates of the first and last gene from these sytenic blocks were used for the construction of the Circos[49] image. The synonymous substitution rates between homologous genes were inferred using Nei Gojobori approach implemented in PAML[44].

7. Genes involved in subterranean fructification, oil biosynthesis and encoding allergens.

7.1 Genes involved in gravitropism and photomorphogenesis

In order to identify the genes involved in gravitropism in *A. duranensis,* a total of 162 genes falling into the GO category "gravitropism" (GO:0009630) and 36 genes identified in *Arabidopsis* were extracted from proteome of *Arabidopsis* and searched against the *A. duranensis* gene set using Blastp with an E-value cutoff of 1e-10. The Blastp hits are then filtered based on 80% query coverage. Of the 198 gravitropism related genes, 137 had homologs in *A. duranensis*. The unidentified gravitropism-related genes is likely due to absence or mis-annotation of the *A. duranensis* genome. Further analysis based on previous functional studies[50-62] identified 24 *A. duranensis* genes likely to be gravitropic including 4 involved in gravity perception, 8 in signal transduction and 12 in organ response (**Dataset S15**). To identify photomorphogenesis-related genes in *A. duranensis,* a total of 280 genes related to photomorphogenesis identified in Arabidopsis were found to have 137 *A. duranensis* homologs using Blastp with an E-value cutoff of 1e-10. The values of Ka and Ks and the ω (Ka/Ks) were estimated between homologous genes using Nei-Gojobori approach implemented in PAML[44].

7.2 Genes involved in oil biosynthesis

Genes involved in oil biosynthesis in *Arabidopsis* (http://aralip.plantbiology.msu.edu/downloads) were retrieved from *Arabidopsis* proteome and searched (BLASTP E-value 1e-5) against soybean and peanut proteomes, independently. The resulting hits obtained from soybean and peanut were then mapped back to the categories as in the aralip database to obtain numbers.

7.3 Allergen-encoding genes

To date, at least 11 potential allergen proteins (Ara h 1-11) have been officially recognized by the International Union of Immunological Societies (IUIS, http://www.allergen.org/Allergen.aspx, last accessed December 12, 2014). These proteins were downloaded from GenBank and subjected to BLASTp analysis against the *A. duranensis* gene set with an E-value cutoff of 1e-30. Of the 11 allergens, nine were found in *A. duranensis*. Of the remained two allergens, the Ara h 6 was identified with an E-value cutoff of 1e-20, the other one (Ara h 4) has been renamed as Ara h 3. All known peanut allergens were identified in *A. duranensis* with an E value cutoff of 1e-20. In order to identify novel allergen-encoding genes in *A. duranensis*, 61 allergen proteins from other crops, like wheat, soybean and tomato, were also downloaded from IUIS. We searched for *A. duranensis* genes orthologous to these allergen-encoding genes, and identified 21 putative orthologs including 13 potential novel allergen-encoding genes as well as 7 orthologs of known peanut allergen genes (**Dataset S16**). For further annotation, these genes were subject to similarity search against the Pfam database (http://pfam.xfam.org/ last accessed December 13, 2014) with an E-value cutoff of 1e-5. These allergen-encoding genes were classified in 14 Pfam families, of which four families contain at least two genes. It is worth to note that Ara h 8 has three paralogs in the *A. duranensis* genome, and the identity between the paralogs ranging from 92~94%.

SI Tables:

Table S1. Construction of libraries, generation and filtering of sequencing data used for genome assembly

Platform	Library	Read Count	Average read length (bp)	Raw data (bp)	Sequence depth
Illumina	250 bp	138,068,824	125	34,517,206,000	25.01
	500 bp	137,054,823	125	34,263,705,750	24.83
	800 bp	115,096,083	125	28,774,020,750	20.85
	2000 bp	71,225,552	125	17,806,388,000	12.90
	5000 bp	91,106,606	125	22,776,651,500	16.50
	10000 bp	61,580,712	125	15,395,178,000	11.16
	20000 bp	305,601,609	125	76,400,402,250	55.36

Table S2. Summary of the *A. duranensis* genome assembly

	Contigs		**Scaffolds**	
	Size	**Number**	**Size**	**Number**
N90	5,864	36,381	148,975	1,718
N80	11,725	24,839	264,326	1,197
N70	17,450	18,084	376,360	864
N60	23,279	13,268	500,641	619
N50	29,584	9,555	649,840	437
Longest (bp)	285,529		5,342,956	
Total size (bp)	972,902,491		1,051,523,805	
Total number (≥ 100 bp)		90,568		8,173
Total number (≥ 1kb)		67,603		5,025
Total number (≥ 2 kb)		54,773		3,996

Table S3. Distribution of contig and scaffold length for *A. duranensis* genome

	Contig				**Scaffold**			
Length (kb)	Number	Average length (bp)	Subtotal length (Mb)	Percentage (%)	Number	Average length (bp)	Subtotal length (Mb)	Percentage (%)
≥100	387	125,938	48.74	5.01	2,084	475,715	991.4	94.28
≥50	3,552	72,807	258.6	26.58	2,544	403,017	1,025	97.50
≥30	9,339	51,402	480.0	49.34	2,799	369,896	1,035	98.46
≥20	15,749	40,468	637.3	65.51	3,003	346,451	1,040	98.94
≥10	27,471	29,373	806.9	82.94	3,356	311,461	1,045	99.40
≥2	54,773	17,190	941.6	96.78	3,996	262,271	1,048	99.67
≥1	67,619	14,197	959.9	98.67	5,027	208,754	1,049	99.80

Table S4. Assessment of the assembled genome through PCR amplification

Category	Fragment number
Total primer pairs used	411
Number of amplified primers	365
Number of non-amplified primers	46
Primers with single amplified fragment	264
Primers with multiple amplified fragments	101
Primers with major amplified fragment	50

Table S5. Evaluation of completeness of the genome assembly using core eukaryotic gene mapping approach (CEGMA)

Parameter		Number	Percent (%)
Total KOGs		458	
One KOG align one gene		410	89.52
One KOG align one gene	overlap>0.7	370	80.78
	overlap >0.5	404	88.21
One KOG align several genes		31	6.76
One KOG align no gene		17	3.71

KOGs=Eukaryotic orthologous gene sequences

Table S6. Assessment of gene space captured in genome assembly using all libraries

	Illumina PE Reads	Illumia MP Reads	454 Reads
Total Reads	781,795,634.00	565,857,140	48,433,168
Mapped reads	688,385,989.00	480,729,955	47,943,412
Mapping percentage (%)	88.05%	84.97%	98.99
Genome coverage at >= 1x (%)	84.68%	63.56%	85.42
Genome coverage at >=2 x (%)	83.00%	60.55%	83.76
Genome coverage at >= 5x (%)	78.14%	54.08	77.65
Genome coverage at >= 10x (%)	70.34%	44.485	63.77
Genome coverage at >= 15x (%)	61.79%	35.875	47.16
Bases not covered (bp)	326,410,883	775,155,378	157,073,491
% of bases not covered	15.15	33.98	14.58
Average depth	31.97	21.56	22.87
Total bases (bp)	2,150,737,583	2,150,841,427	1,077,216,168

Table S7. Estimation of *A. duranensis* genome using K-mer statistics

K-mer value	K-mer number	Depth	Genome size (bp)	Used bases	Used reads	Depth (X)	Average read length (bp)
25	60,198,113,206	24	1,381,794,909	78,961,359,034	781,793,678	57.14	101

Table S8. Gene sets used in this study from different plant species

Species	Database	Version
Soybean	Phytozomev9.1	JGI Glyma1.1 annotation of the chromosome-based Glyma1 assembly
Medicago	Phytozomev9.1	Mt3.5v4 on assembly MedtrA17_3.5 from the Medicago Genome Sequence Consortium
Lotus	kazusa.or.jp	lotus_r2.5
Common bean	Phytozomev9.1	JGI annotation v1.0 on assembly v1.0 using published ESTs, and JGI RNAseq
Chickpea	Legume Information System	v1.0
Pigeonpea	Legume Information System	v1.0
Canola	Phytozomev9.1	BrapaFPsc_277_v1.3
Cotton	Phytozomev9.1	JGI annotation v2.1 on assembly v2.0
Castor	Phytozomev9.1	TIGR release 0.1
Linseed	Phytozomev9.1	Lusitatissimum_200_v1.0
Arabidopsis	Phytozomev9.1	TAIR release 10 acquired from TAIR
Apple	Phytozomev9.1	GDR prediction v1.0 on Malus x domestica assembly v1.0
Poplar	Phytozomev9.1	JGI assembly release v3.0, annotation v3.0
Tomato	Phytozomev9.1	SGNTomato Genome Project ITAG2.3
Rice	Phytozomev9.1	MSU Release 7.0 of the Rice Genome Annotation

Table S9. The statistics of aligned genes between *A. duranensis* and other plant species with an E value cutoff of 1e-5.

Species	*A. duranensis*		Aligned species	
	Matched genes	Percentage	Matched genes	Percentage
A. duranensis vs Arabidopsis	22132	43.98	13185	48.09
A. duranensis vs Canola	23129	45.96	13973	34.51
A. duranensis vs Chickpea	24496	48.68	14116	49.93
A. duranensis vs Pigeonpea	31456	62.51	16570	34.04
A. duranensis vs Soybean	26836	53.33	17445	31.13
A. duranensis vs Cotton	22990	45.68	15733	41.95
A. duranensis vs Lotus	27400	54.45	13079	33.99
A. duranensis vs Linseed	21798	43.32	14082	32.39
A. duranensis vs Apple	23827	47.35	14468	22.78
A. duranensis vs Medicago	28831	57.29	16081	31.60
A. duranensis vs Rice	22572	44.85	12224	30.07
A. duranensis vs Poplar	23899	47.49	15840	38.32
A. duranensis vs Common bean	26978	53.61	15391	56.59
A. duranensis vs Castor	25245	50.16	13689	43.85
A. duranensis vs Tomato	23640	46.98	13205	38.03

Table S10. General statistics of predicted protein-coding genes in *A. duranensis and* comparison with other plant species

Gene set	Common name	Number of genes	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average intron length (bp)
Reference	*A. duranensis*	50,324	3,057.92	312.36	3.37	709.57
Homology	Soybean	56,044	4,671.51	214.91	10.22	486.85
	Medicago	50,894	3,064.99	231.70	5.87	413.13
	Lotus	38,482	1,494.66	258.73	2.96	447.89
	Common bean	27,197	4,048.62	234.25	6.85	449.59
	Chickpea	28,269	3,055.39	236.51	4.93	448.78
	Pigeonpea	48,680	2,348.70	267.39	3.59	458.45
	Canola	40,492	2,274.32	230.81	5.63	185.02
	Cotton	37,505	3,914.53	203.66	14.07	333.79
	Castor	31,221	2,261.54	242.46	4.17	339.80
	Linseed	43,471	2,307.97	238.58	5.03	260.97
	Arabidopsis	27,416	2,335.51	220.87	7.57	150.06
	Apple	63,514	2,639.37	236.13	4.74	383.85
	Poplar	41,335	3,759.28	211.44	11.13	359.97
	Tomato	34,727	3,163.56	228.78	4.62	505.12
	Rice	40,648	3,169.63	240.49	5.90	370.11

[#]Protein sequences from 15 sequenced plant species were used to perform gene prediction, taking one species each time. We mapped them to the genome assembly using TblastN (E-value- 1e-5). After this, homologous genome sequences were aligned against the matching proteins for accurate spliced alignments.

Table S11. Functional annotation of predicted genes in *A. duranensis*

Database	Number	Percentage
SWISS-PROT	20,701	41.13 %
TrEMBL	35,365	70.27 %
NR	35,726	70.99%
NT	40,552	80.58%
InterPro	30,032	59.68 %
KEGG	30,573	60.75 %
GO	24,498	48.68 %
Pfam	25,771	51.21 %
CDD	23,903	47.50 %
Un-annotated	5,494	10.9%

Table S12. Details on gene family for *A. duranensis* and other plant species

Species	Total predicted genes	Genes in orthologous groups	Genes not in orthologous groups [1]	Total orthologous groups[2]	Species-specific homolog groups[3]	Average genes group
A. duranensis	50,324	40,736	9,588	14,005	1,423 (16,472)	2.91
Chickpea	31,988	30,412	1,576	14,657	348 (1,375)	2.07
Pigeonpea	48,680	42,353	6,327	17,222	1,440 (7,934)	2.46
Soybean	73,320	62,797	10,523	17,900	1,265 (3,331)	3.51
Medicago	45,888	32,786	13,102	14,159	2,202 (9,533)	2.32
Lotus	42,399	24,345	18,054	14,599	1,155 (4,248)	1.67
Common bean	31,638	29,666	1,972	15,908	271 (801)	1.86
Linseed	43,484	37,033	6,451	14,258	1,474 (4,907)	2.60
Canola	101,040	81,965	19,075	21,752	5,582 (17,934)	3.77
Castor	31,221	21,077	10,144	15,360	604 (1,608)	1.37
Cotton	77,267	71,534	5,733	16,910	2,016 (6,652)	4.23
Rice	49,061	36,506	12,555	13,995	2,900 (10,795)	2.61
Tomato	34,727	26,231	8,496	14,260	983 (3,873)	1.84
Apple	63,517	48,160	15,357	17,200	3,659 (11,308)	2.80
Arabidopsis	35,386	31,882	3,504	16,329	577 (1,550)	1.95
Poplar	73,013	64,680	8,333	16,465	1,475 (4,410)	3.93

[1] Predicted genes that were not organized into groups using OrthoMCL. We suggest that many such genes are misannotated, though we cannot rule out genes with unique domain arrangements that have undergone lineage specific expansion. [2] Orthologous groups containing at least one gene from the indicated species. [3] Groups containing putative paralogs from the indicated species, but lacking genes from other species. Such unassigned homologous groups may contain genes with ambiguous relationships among species, such as many of the NBS-LRR disease resistance genes that can evolve by processes such as non-allelic recombination and gene conversion.

Table S13. Comparison of *A. duranensis* gene families with soybean and *Medicago*

Family	**Difference in gene copy number**															
size	**<-6**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**>6**	**Total**
Shared gene families between *A. duranensis* and Soybean																
1								94.66	4.49	0.31	0.08	0.08	0.02	0.36	0.34	6197
2							85.43	12.9	1.07	0.25	0.13	0.03	0.03	0.09	0.06	3170
3						67.87	25.11	5.48	0.88	0	0	0	0.11	0.22	0.22	912
4					48.89	31.67	15.56	2.5	0	0	0	0	0	0.28	0.28	360
5				34.01	34.52	19.29	6.09	2.54	0	0.51	0	0	0	0.51	0.51	197
6			33.66	34.65	17.82	6.93	0.99	0	0	0	0	0	0	0	0	101
7		8.57	31.43	25.71	11.43	10	2.86	0	0	0	0	0	0	0	0	70
8	19.35	12.9	16.13	11.29	9.68	9.68	1.61	1.61	1.61	1.61	0	0	0	1.61	1.61	62
9	22.73	25	13.64	6.82	6.82	2.27	2.27	0	0	0	0	0	0	0	0	44
10	33.33	2.56	15.38	12.82	2.56	5.13	0	2.56	0	0	0	0	0	0	0	39
Shared gene families between *A. duranensis* and *Medicago*																
1								90.34	8.33	0.77	0.22	0.08	0	0.24	0.22	6325
2							67.24	24.14	6.65	0.99	0.25	0	0	0.57	0.57	1218
3						49.86	24.23	16.06	4.23	2.25	1.13	0.28	0.28	0.85	0.85	355
4					30.67	28	20	10	4.67	1.33	1.33	0	0	1.33	1.33	150
5				33.72	16.28	16.28	9.3	5.81	5.81	2.33	1.16	1.16	0	2.33	1.16	86
6			26.79	16.07	17.86	10.71	8.93	3.57	1.79	3.57	0	0	0	0	0	56
7		16.67	9.52	7.14	9.52	11.9	11.9	9.52	4.76	2.38	0	0	0	0	0	42
8	16.67	23.33	6.67	3.33	6.67	6.67	0	3.33	0	0	3.33	0	0	3.33	3.33	30
9	37.04	3.7	7.41	3.7	0	3.7	3.7	3.7	0	0	3.7	0	0	0	0	27
10	25	3.57	3.57	7.14	0	3.57	3.57	3.57	7.14	0	3.57	3.57	0	0	0	28

Table S14. Details on single copy orthologs and unique paralogs in *A. duranensis* and other plant species

Species	Single-copy orthologs	Co-orthologs[1]	Unique paralogs	Other orthologs[2]	Unclustered genes
A. duranensis	9,968	7,138	16,472	7,158	9,588
Chickpea	8,346	12,125	1,375	8,566	1,576
Pigeonpea	11,022	11,201	7,934	12,196	6,327
Soybean	3,779	25,189	3,331	30,498	10,523
Medicago	7,929	9,403	9,533	5,921	13,102
Lotus	10,120	7,931	4,248	2,046	18,054
Common bean	9,791	11,609	801	7,465	1,972
Linseed	2,895	14,369	4,907	14,862	6,451
Canola	1,531	25,579	17,934	36,921	19,075
Castor	12,258	7,400	1,608	189	10,144
Cotton	4,808	29,975	6,652	30,099	5,733
Rice	5,654	12,793	10,795	7,264	12,555
Tomato	9,247	9,151	3,873	3,960	8,496
Apple	4,710	15,529	11,308	16,613	15,357
Arabidopsis	9,293	11,947	1,550	9,092	3,504
Poplar	3,674	26,347	4,410	30,249	8,333

[1] Co-orthologous genes, also known as “in-paralogs”, are derived from duplication in the indicated genome. [2] Other orthologs represent gene duplication events internal to the overall set, but basal more than two of the compared species.

Table S15. Summary of predicted non-protein coding genes in *A. duranensis* genome

Type	Sub-type	Number	Average length (bp)	Total length (bp)	Percent (%)
miRNA		816	107.34	87,598	0.0063
tRNA		913	73.28	66,904	0.0048
rRNA	5S rRNA	61	116.59	7,112	0.0005
	5.8S rRNA	17	152.94	2,600	0.0002
	18S rRNA	21	1944.67	40,838	0.0029
	28S rRNA	16	4579.94	73.279	0.0053
	Total rRNA	115	1076.77	123,829	0.0089
snRNA	CD- box snRNA	71	106.73	7,578	0.0005
	Splicing snRNA	131	137.85	18,058	0.0013
	Total snRNA	202	126.91	25,636	0.0018

Table S16. New miRNAs identified in the *A. duranensis* genome

ID	Sequence	Length	Scaffold	Start	End	Strand
Peanut_m0002-3p	ATAACCAAGGAAAAGACATT	20	scaffold1297	106541	106560	-
Peanut_m0003-3p	ACTTAGGCCTTAGAACTTAT	20	scaffold18250	900	919	+
Peanut_m0004-3p	ACATTAAACATGGGACAATTTA	22	scaffold1988	30519	30540	+
Peanut_m0007-3p	TGAGATATCTCTTCCAGAAG	20	scaffold371	58057	58076	-
Peanut_m0009-3p	GACTGTAGAGTGGTAATTCAA	21	scaffold426	160999	161019	-
Peanut_m0014-3p	ACAGCCATTTTTGCCGAGTT	20	scaffold918	204369	204388	-
Peanut_m0001-5p	CAGGAGACCCGGGTTCGATTCCC	23	scaffold1221	110014	110036	+
Peanut_m0005-5p	CTTTAGGTCAATGATTGGTA	20	scaffold2433	93926	93945	-
Peanut_m0006-5p	AGTTCTGAGAAGTCTTCTTTG	21	scaffold3536	27272	27292	-
Peanut_m0008-5p	AGAAGAACTTGTAGGTGTTGAA	22	scaffold4210	29738	29759	-
Peanut_m0010-5p	GAGGAGACAGAAACAGGTAG	20	scaffold454	183899	183918	-
Peanut_m0011-5p	TGACTTTTGGAAAATGTTTG	20	scaffold495	204813	204832	+
Peanut_m0012-5p	TTCTGACTTCTTTAGGCAGT	20	scaffold6457	39441	39460	+
Peanut_m0013-5p	TCTCTGCAGAAGGAATGACA	20	scaffold681	121285	121304	-
Peanut_m0015-5p	GTGCAGGACGATGTCGTTGC	20	scaffold9422	15413	15432	+

Table S17. Target genes and their function annotation of new miRNAs in *A. duranensis*

miRNA ID	Number of target genes	CDD (Conserved Domains Database)	Putative functions of target genes
m0001-5p	3	COG1691	NCAIR mutase (PurE)-related proteins
		PLN03195	fatty acid omega-hydroxylase
		TIGR03225	benzoyl-CoA oxygenase, B subunit
m0002-3p	17	cd00303	Retropepsins
		cd11236	MET-like receptor tyrosine kinases
		COG3083	Predicted hydrolase of alkaline phosphatase superfamily
		COG4036	Predicted membrane protein
		COG5222	Uncharacterized conserved protein, contains RING Zn-finger
		pfam04900	Fcf1
		PRK00232	4-hydroxythreonine-4-phosphate dehydrogenase
		PRK06599	DNA topoisomerase I
		PRK08377	NADH dehydrogenase subunit N
		PRK09330	cell division protein FtsZ
		PRK09629	bifunctional thiosulfate sulfurtransferase
		PRK12679	transcriptional regulator Cbl
		PRK13902	alanyl-tRNA synthetase
		TIGR04055	putative heme d1 biosynthesis radical SAM protein NirJ2
m0003-3p	1	TIGR01160	translation initiation factor SUI1, eukaryotic
m0004-3p	1	COG0061	NAD kinase
m0005-5p	1	PLN02393	leucoanthocyanidin dioxygenase like protein
m0006-5p	6	pfam09773	Meckelin (Transmembrane protein 67)

		pfam13639	Ring finger domain
		PRK13897	type IV secretion system component VirD4
		PTZ00350	adenylosuccinate synthetase
		smart00220	Serine/Threonine protein kinases, catalytic domain
m0007-3p	1	pfam03124	EXS family
m0008-5p	9	cd00180	Catalytic domain of Protein Kinases
		pfam05133	Phage portal protein, SPP1 Gp6-like
		pfam05297	Herpesvirus latent membrane protein 1 (LMP1)
		pfam06291	Bor protein
		PLN02499	glycerol-3-phosphate acyltransferase
		PRK04028	glutamyl-tRNA(Gln) amidotransferase subunit E
		PTZ00479	RAP Superfamily
		TIGR02168	chromosome segregation protein SMC, common bacterial type
		TIGR03981	His-Xaa-Ser system putative quinone modification maturase
m0010-5p	1	PLN02311	chalcone isomerase
m0011-5p	6	cd01851	Guanylate-binding protein (GBP) family (N-terminal domain)
		cd08866	Ligand-binding SRPBCC domain
		pfam00587	tRNA synthetase class II core domain (G, H, P, S and T)
		PHA02746	protein tyrosine phosphatase
		PLN03240	putative Low-temperature-induced protein
m0013-5p	2	pfam05699	hAT family dimerisation domain
		TIGR02169	chromosome segregation protein SMC
m0015-5p	1	COG1752	Predicted esterase of the alpha-beta hydrolase superfamily

Table S18. Summary of simple sequence repeats in *A. duranensis* regarding their distribution and primer design for peanut genetics and breeding applications.

SSR Statistics	**Numbers**
Total number of sequences examined	20,597
Total size of examined sequences (bp)	1,077,216,168
Total number of identified SSRs	105,003
Number of SSR containing sequences	15,209
Number of sequences containing more than 1 SSR	12,308
Number of SSRs present in compound formation	25,672
Distribution to different repeat type classes	
Number of di-nucleotide repeats	45,622
Number of tri-nucleotide repeats	32,070
Number of tetra-nucleotide repeats	3,966
Number of penta-nucleotide repeats	1,450
Number of hexa-nucleotide repeats	428
Number of compound repeats	21,467
Primer pairs for SSRs	
Scaffolds were used to design primer pairs	11,712
Total numbers of primer pairs designed	84,464

Table S19. Details on re-sequencing data of ten genotypes including four synthetic tetraploids and six diploids

Germplasm	Ploidy (genome)	Parental combinations	Read type	Number of reads	Read length (bp)	Data size (bp)
ISATGR_5	Synthetic tetraploid (BB)	[*A. magna* (ICG 8960) x *A. batizocoi* (ICG 8209)]	Paired end	983,446,602	101	99,328,106,802
ISATGR_278-18	Synthetic tetraploid (AB)	[*A. duranensis* (ICG 8138) x *A. batizocoi* (ICG 13160)]	Paired end	1,230,617,008	101	124,292,317,808
ISATGR_1212	Synthetic tetraploid (AB)	[*A. duranensis* (ICG 8123) x *A. ipaensis* (ICG 8206)]	Paired end	914,091,908	101	92,323,282,708
ISATGR_184	Synthetic tetraploid (AB)	[*A. ipaensis* (ICG 8206) x *A. duranensis* (ICG 8123)]	Paired end	1,258,898,410	101	127,148,739,410
ICG_8123	Diploid (A)	*A. duranensis*	Paired end	504,473,764	101	50,951,850,164
ICG_8138	Diploid (A)	*A. duranensis*	Paired end	503,836,436	101	50,887,480,036
ICG_8960	Diploid (B)	*A. magna*	Paired end	461,986,170	101	46,660,603,170
ICG_8209	Diploid (B)	*A. batizocoi*	Paired end	458,037,642	101	46,261,801,842
ICG_13160	Diploid (B)	*A. batizocoi*	Paired end	487,100,820	101	49,197,182,820
ICG_8206	Diploid (B)	*A. ipaensis*	Paired end	553,199,484	101	55,873,147,884

Table S20: Distribution of SNPs identified among the A genomes (two genotypes) and B genomes (four genotypes)

	ICG 8123		ICG 8138		ICG_8960		ICG_8209		ICG_13160		ICG_8206	
	SNP	Rate	SNP	Rate	SNP	Rate	SNP	Rate	SNP	Rate	SNP	Rate
Gene region	1,437,202	16.677	1,438,084	16.618	1,243,501	33.479	1,280,304	32.964	1,262,785	34.271	1,274,376	32.868
Exon	453,740	5.265	450,314	5.204	423,103	11.391	429,687	11.063	428,290	11.623	436,853	11.267
Intron	968,333	11.237	972,585	11.239	807,968	21.753	838,142	21.579	822,277	22.316	824,558	21.266
Others	15,129	0.176	15,185	0.175	12,430	0.335	12,475	0.321	12,218	0.332	12,965	0.334
ncRNA	1,567	0.018	1,311	0.015	849	0.023	896	0.023	885	0.024	771	0.02
tRNA	253	0.003	229	0.003	71	0.002	67	0.002	81	0.002	55	0.001
rRNA	407	0.005	202	0.002	110	0.003	125	0.003	131	0.004	101	0.003
snRNA	248	0.003	240	0.003	153	0.004	190	0.005	186	0.005	146	0.004
miRNA	659	0.008	640	0.007	515	0.014	514	0.013	487	0.013	469	0.012
TEs	792,959	9.201	790,347	9.133	219,683	5.915	217,558	5.601	203,212	5.515	230,178	5.937
Others	6,385,994	74.103	6,424,066	74.234	2,250,204	60.583	2,385,247	61.412	2,217,848	60.19	2,371,974	61.176
Total	8,617,722	100	8,653,808	100	3,714,237	100	3,884,005	100	3,684,730	100	3,877,299	100

Table S21: Summary of structural variations in diploid (A-genome and B-genome) and synthetic (AB-genome) genotypes

	No. of SVs		**Total length (kb)**	**Average length (bp)**
Diploid A genome				
Sample	**PI 475845-reference genome**			
Insertion	0		0	0
Deletion	33,648		116,094.635	3,450.269
Inversion	3,003		55,763.004	18,569.099
CNVs	15,958	gain : 4,243	-	-
		loss : 11,715		
Sample	***ICG 8138***			
Insertion	0		0	0
Deletion	23,077		122,149.219	5,293.115
Inversion	2,234		46,606.732	20,862.458
CNVs	20,776	gain : 11,858	-	-
		loss : 8,918		
Sample	***ICG 8123***			
Insertion	0		0	0
Deletion	22,600		119,789.414	5,300.417
Inversion	2,084		43,681.400	20,960.365
CNVs	20,955	gain : 12,369	-	-
		loss : 8,586		
Diploid A genome				

Sample	***ICG 8960***			
Insertion	0		0	0
Deletion	8,946		60,149.980	6,723.673
Inversion	1,378		32,379.408	23,497.393
CNVs	24,132	gain : 13,288	-	-
		loss : 10,844		
Sample	***ICG 8209***			
Insertion	0		0	0
Deletion	9,723		61,424.708	6,317.465
Inversion	1,417		29,805.699	21,034.368
CNVs	24,146	gain : 10,729	-	-
		loss : 13,417		
Sample	***ICG 13160***			
Insertion	0		0	0
Deletion	10,344		61,214.867	5,917.911
Inversion	1,396		32,590.761	23,345.817
CNVs	24,188	gain : 9,559	-	-
		loss : 14,629		
Sample	***ICG 8206***			
Insertion	0		0	0
Deletion	9,801		64,806.025	6,612.185
Inversion	1,488		36,921.970	24,813.152
CNVs	24,092	gain : 11,957	-	-
		loss : 12,135		

Synthetic genotypes				
Sample	***ISATGR-5***			
Insertion	0		0	0
Deletion	15,149		87,538.429	5,778.496
Inversion	2,380		53,421.757	22,446.116
CNVs	24,434	gain : 18,733	-	-
		loss : 5,701		
Sample	***ISATGR 278-18***			
Insertion	0		0	0
Deletion	29,842		158,257.555	5,303.182
Inversion	3,662		76,232.457	20,817.165
CNVs	20,913	gain : 13,367,	-	-
		loss : 7,546		
Sample	***ISATGR 1212***			
Insertion	0		0	0
Deletion	24,747		135,350.592	5,469.374
Inversion	3,184		68,492.371	21,511.423
CNVs	20,939	gain : 13,219	-	-
		loss : 7,720		
Sample	***ISATGR 184***			
Insertion	1		0.217	217
Deletion	31,651		168,895.231	5,336.174
Inversion	3,802		79,680.365	20,957.487
CNVs	20,601	gain : 12,384	-	-

Table S22. Summary of putative acyl lipid genes in *A. duranensis*, *Arabidopsis* and soybean

Category of lipid genes	*A. duranensis*	Arabidopsis	Soybean
Phospholipase	115	90	120
Miscellaneous lipid synthesis related	92	73	93
Sphingolipid synthesis	40	28	40
Phospholipid synthesis in mitochondria	16	10	16
Fatty acid synthesis in plastids	73	48	72
Aromatic suberin synthesis	14	8	14
Lipid signaling	187	142	191
Aliphatic suberin synthesis	42	34	42
Eukaryotic phospholipid synthesis	75	45	75
Lipase	330	269	336
Lipid trafficking	10	6	10
Cuticular wax synthesis	191	167	200
Mitochondrial fatty and lipoic acid synthesis	22	13	22
TAG degradation	47	35	47
TAG synthesis	96	68	96
Fatty acid elongation and cuticular wax synthesis	30	26	30
GDSL	127	106	127
Beta-oxidation	35	25	35
Lipid acylhydrolase	15	11	15
Galactolipid degradation	10	7	10
Cutin synthesis	31	28	31
Plastidial glycerolipid, galactolipid and sulfolipid synthesis	73	52	73
Total	1671	1291	1695

Table S23. Summary of samples collected during seed development in peanut

Stages	Samples	Seed size (mm)
P5	Seed	1.0 – 2.0
P6	Seed	2.0 – 4.0
P7	Seed	4.0 – 6.0
P8	Seed	6.0 – 8.0
P9	Seed	8.0 – 10.0
P10	Seed	10.0 – 12.0

SI Figures:

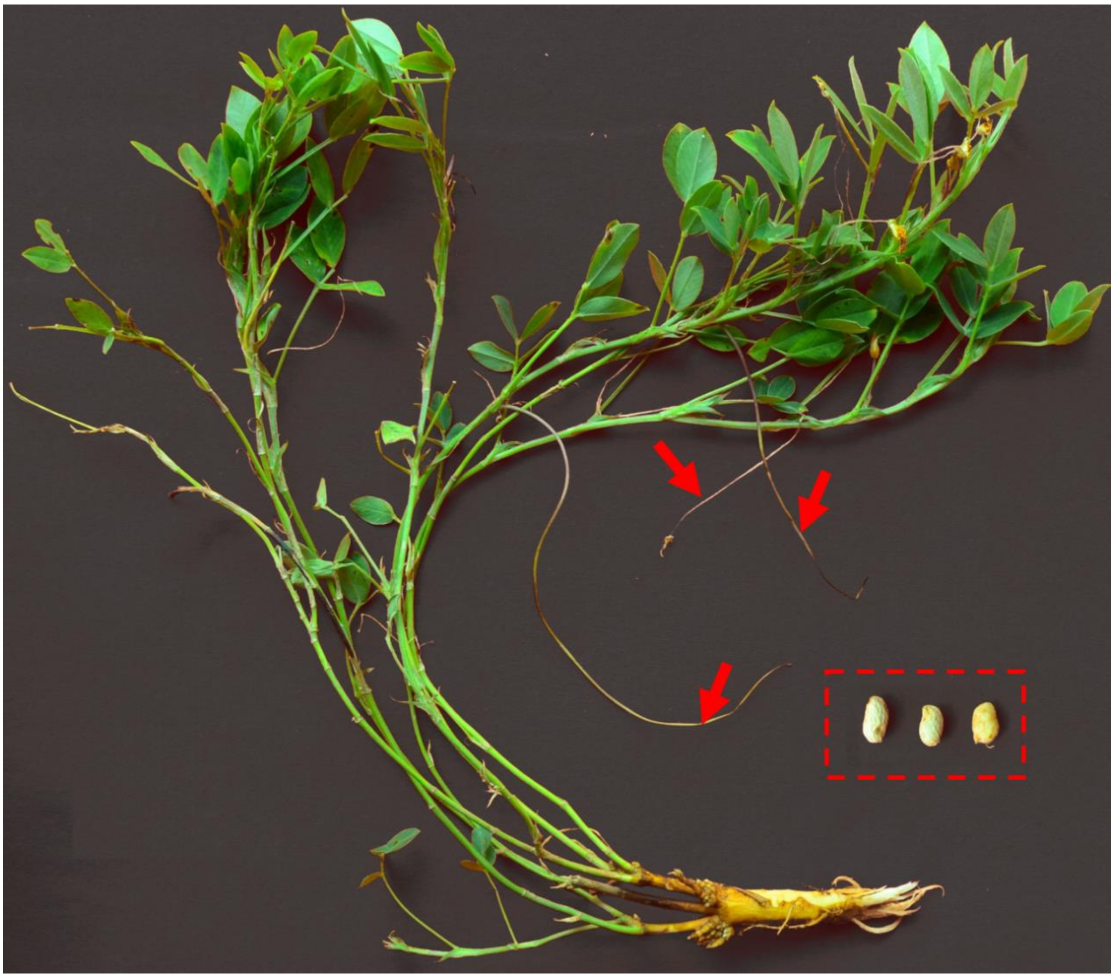


Figure S1. ***A. duranensis* accessions PI475845 (reference genome).** The red arrows show the aerially developing pegs, and the red dash box shows the pods developed underground. Aerially pegs do not normally expand until penetration into the soil. This accession was collected from Tariji Bolivia (Latitude: 21.53, Longitude: 63.38) in 1977 by collectors GKBSPSc (Gregory, W.C.; Krapovickas, A.; Banks, D.J.; Simpson, C.E.; Pietrarelli, J.; and Schinini, A.) (Stalker et al., 1995).

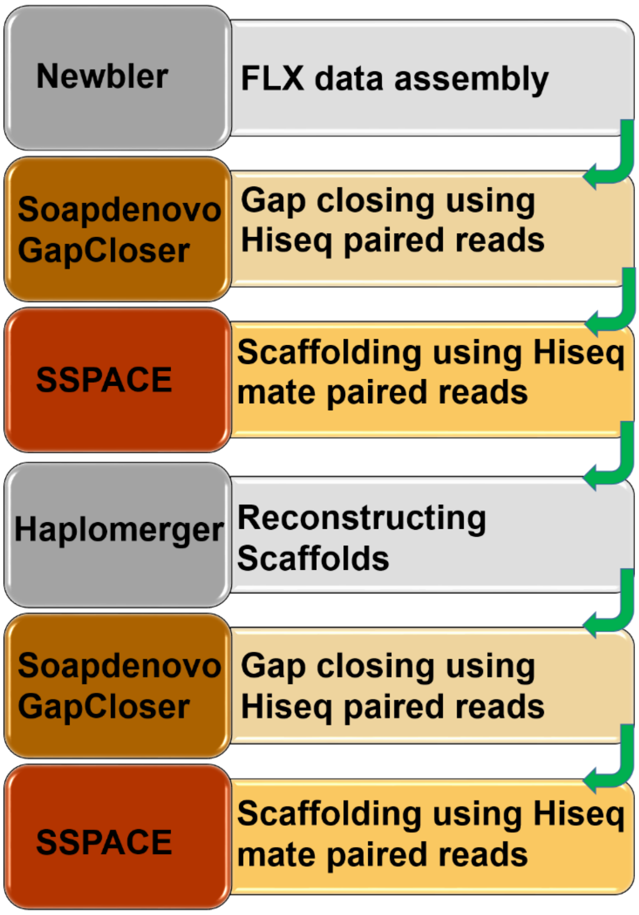

Newbler
FLX data assembly
Soapdenovo GapCloser
Gap closing using Hiseq paired reads
SSPACE
Scaffolding using Hiseq mate paired reads
Haplomerger
Reconstructing Scaffolds
Soapdenovo GapCloser
Gap closing using Hiseq paired reads
SSPACE
Scaffolding using Hiseq mate paired reads

Figure S2. Flowchart of the approaches used for *de novo* assembly

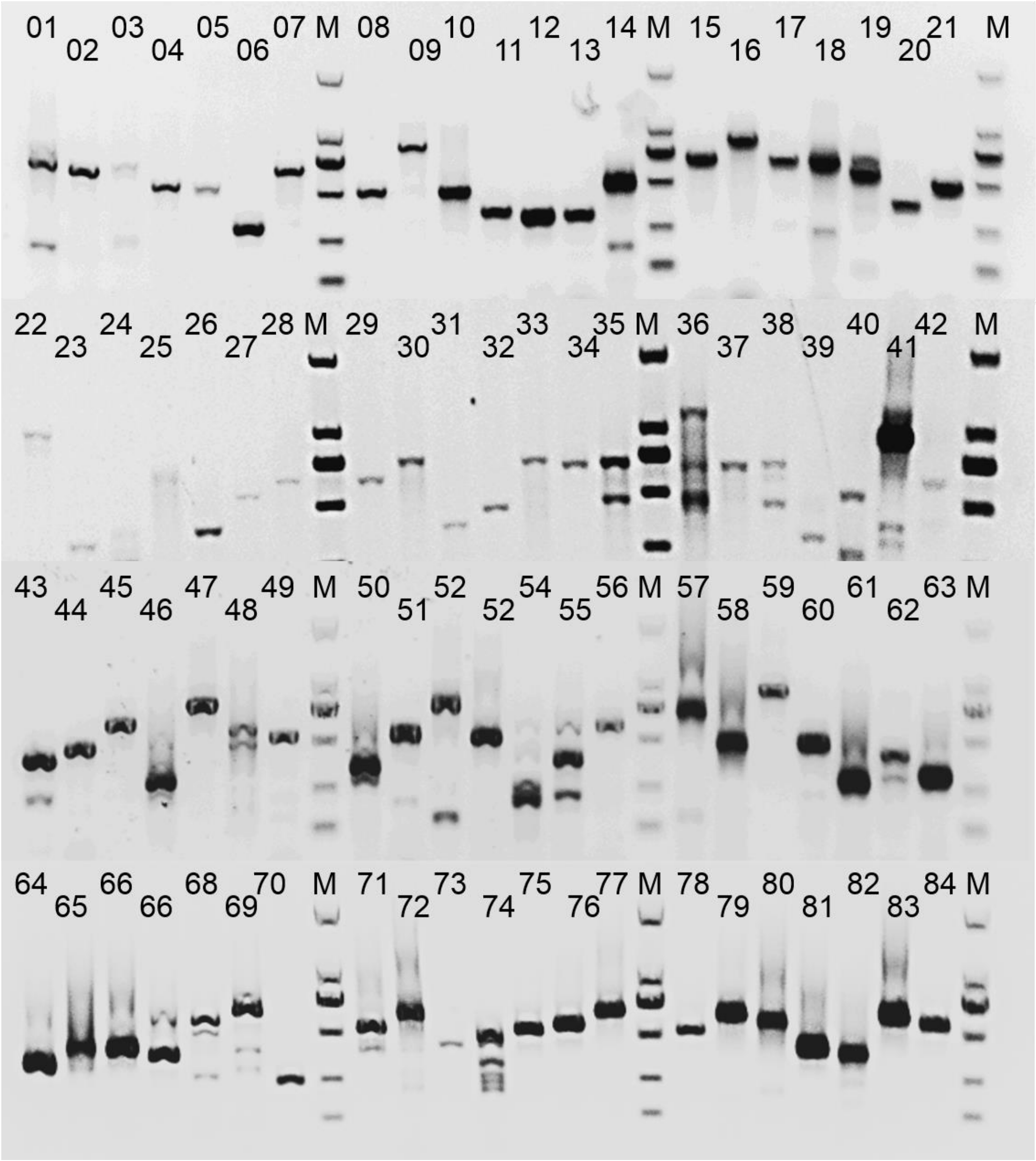


Figure S3. Evaluation of the *A. duranensis* assembled genome using PCR amplification

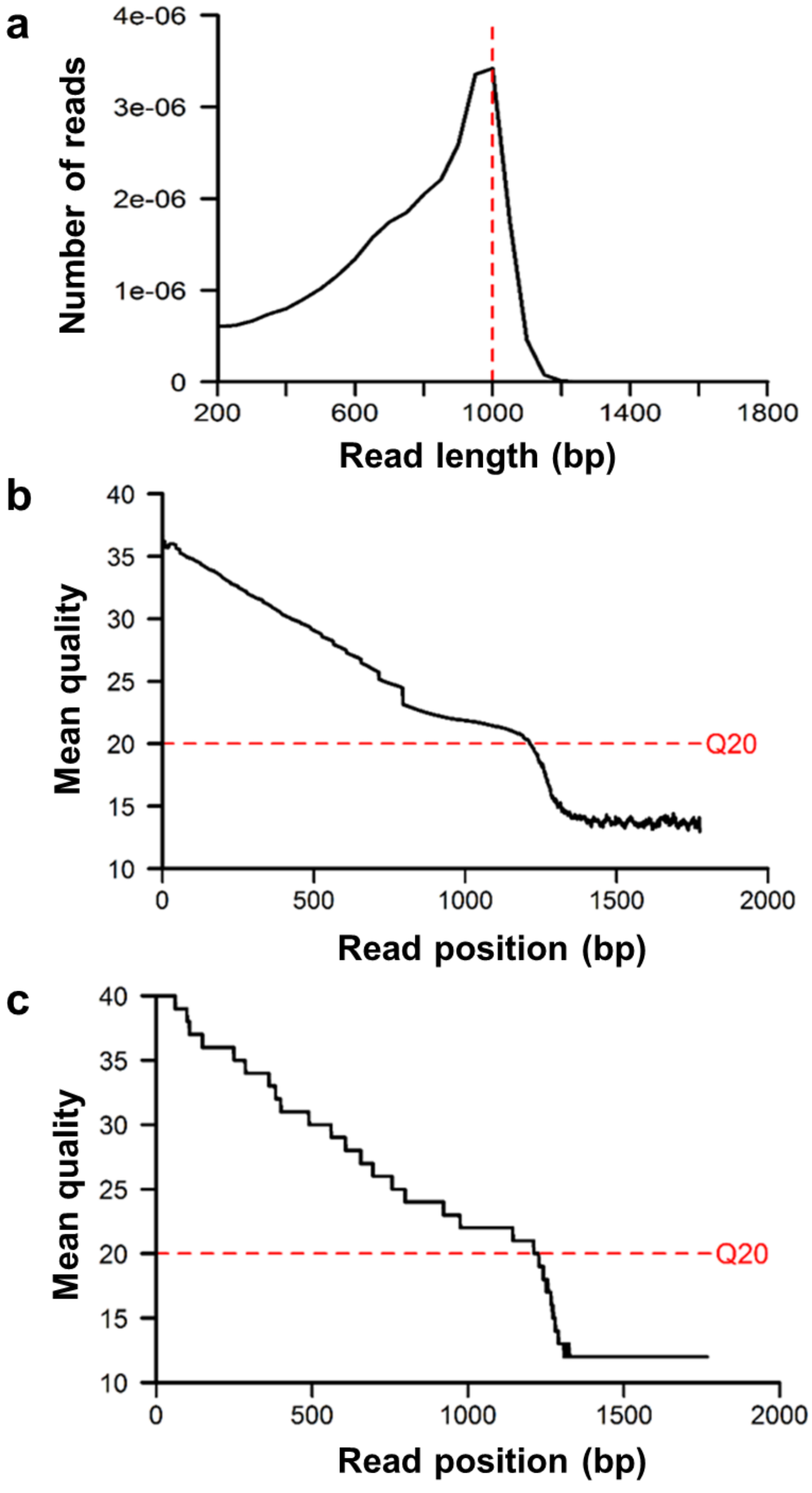


Figure S4. Quality assessment of the sequencing data.

The distributions were computed using FastQC (a) read length distribution, (b) mean read quality per read position, (c) median read quality per read position.

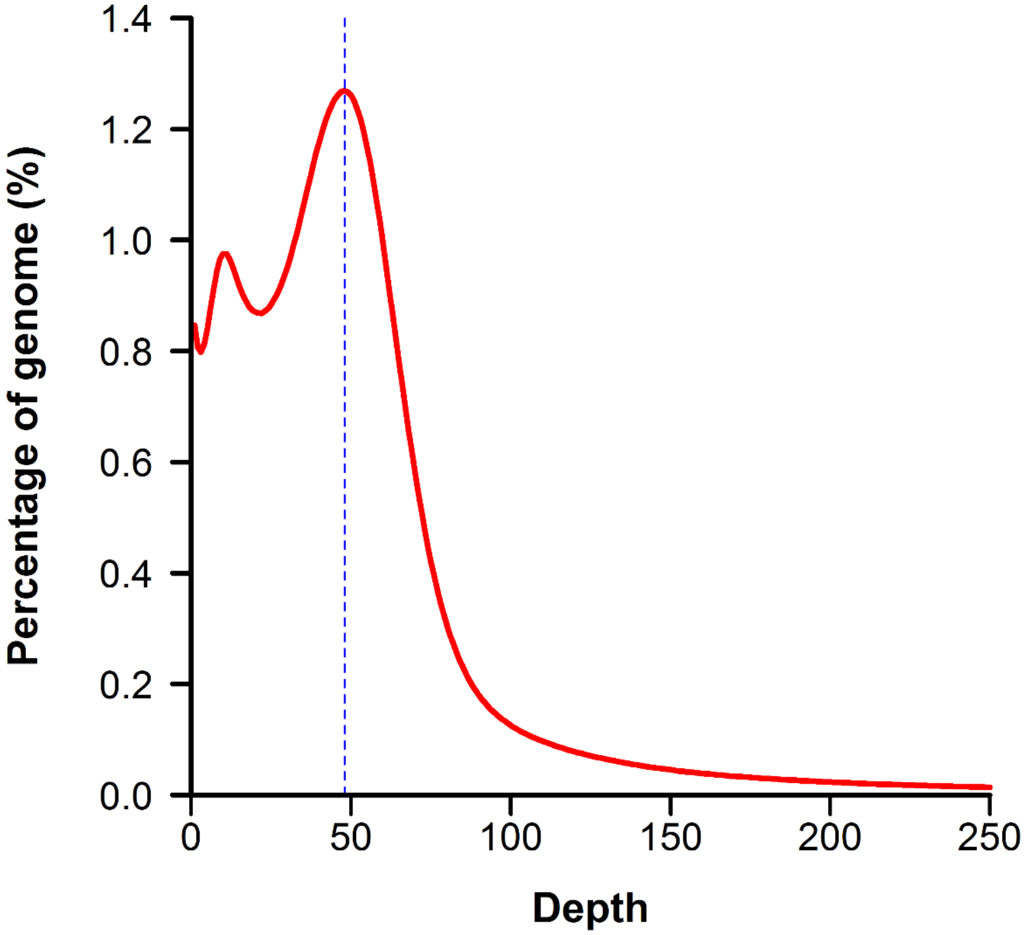


Figure S5. Distribution of sequence depth across the assembled genome.
The Y-axis represents the proportion of the genome at a given sequencing depth.

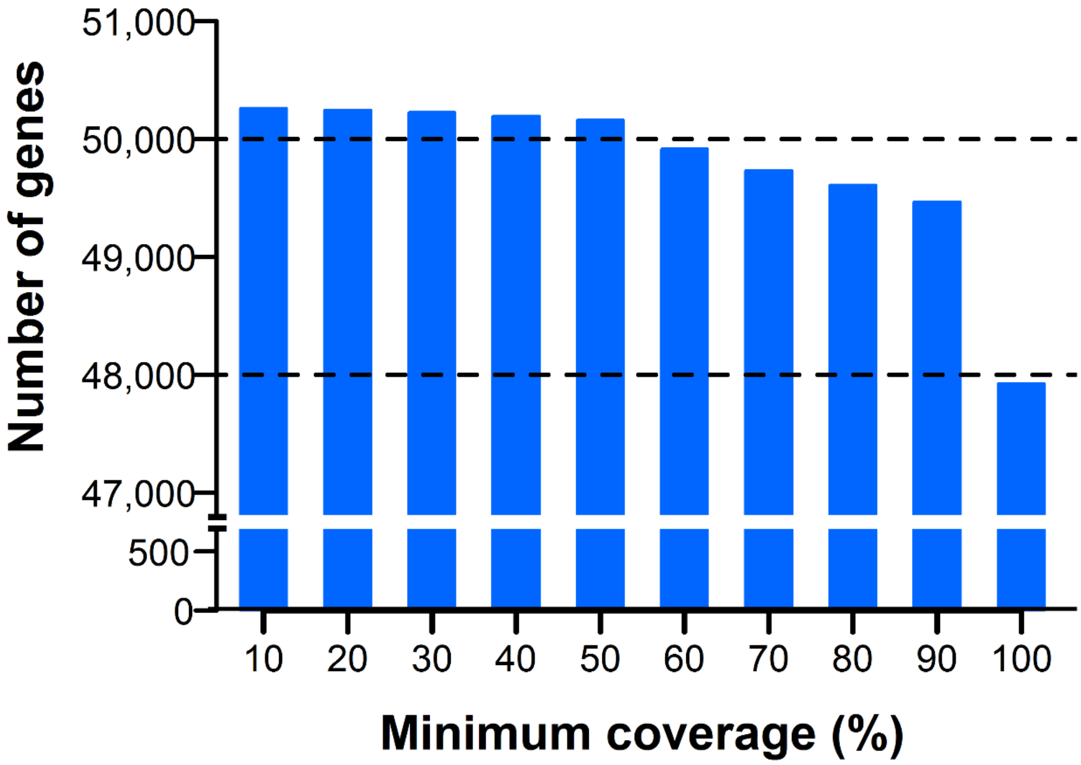


Figure S6. Coverage of transcripts in the *A. duranensis de novo* assembly.
The predicted genes were covered by transcripts with > 98% identity, and the genes in each coverage were counted ranging from 10% to 100%.

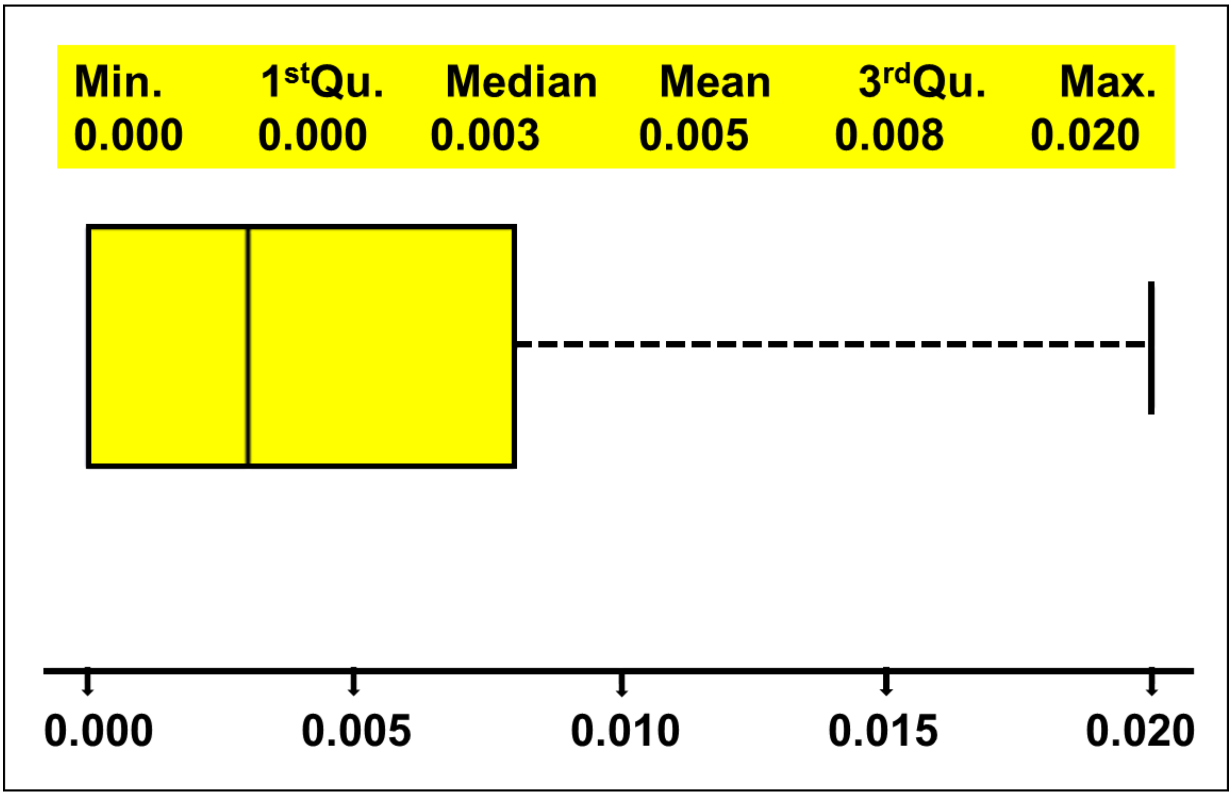


Figure S7. Boxplot of the heterozygosity in 1-kb window of *A. duranensis* genome. Heterozygosity in each of 1 kb window was computed and plotted.
The computed heterozygosity matches well with that by AllpathLG (~3 SNPs per kb).

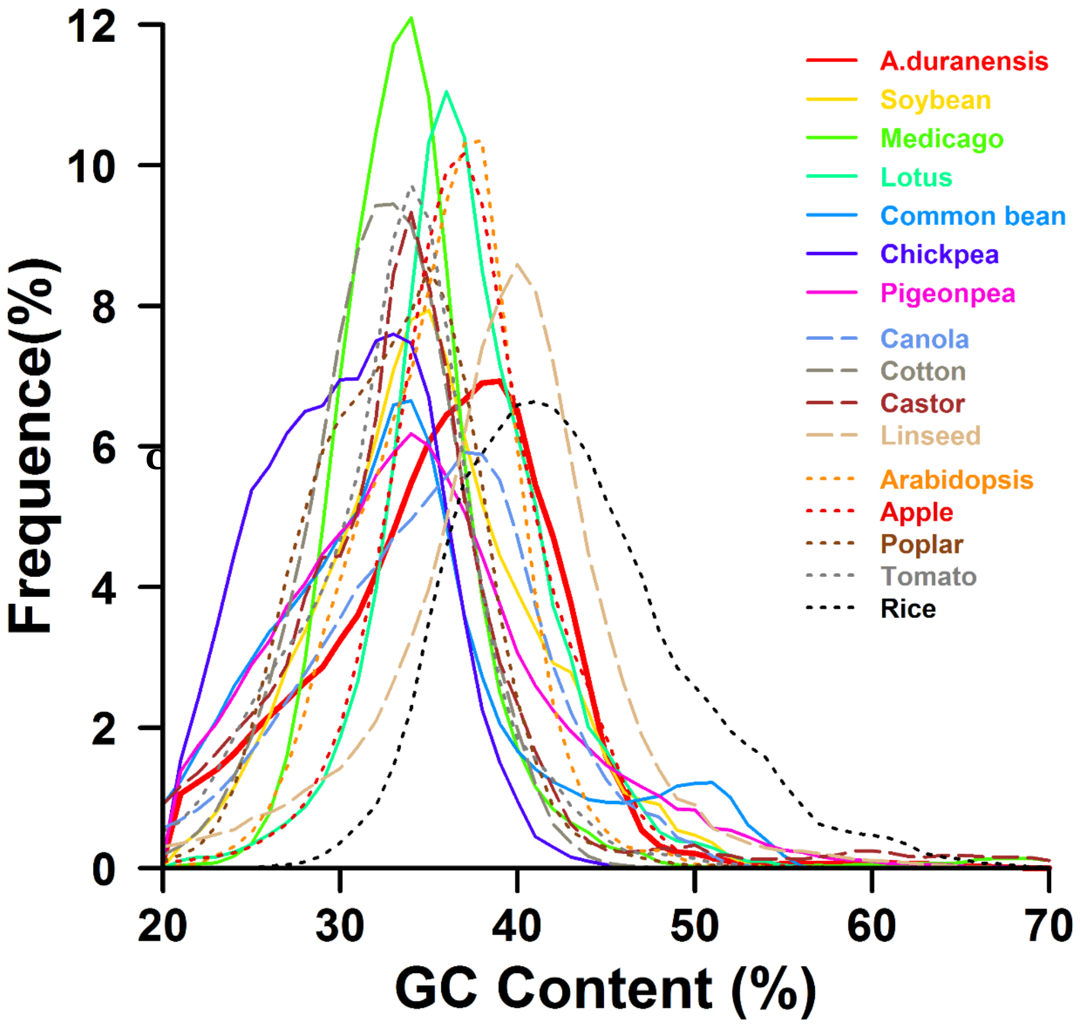


Figure S8. Comparison of GC content distribution and variation among *A. duranensis*, legumes, oilseeds and other plant species.

Solid lines represent legume species, dash lines represent oilseed species, and dot lines represent other distantly related plant species

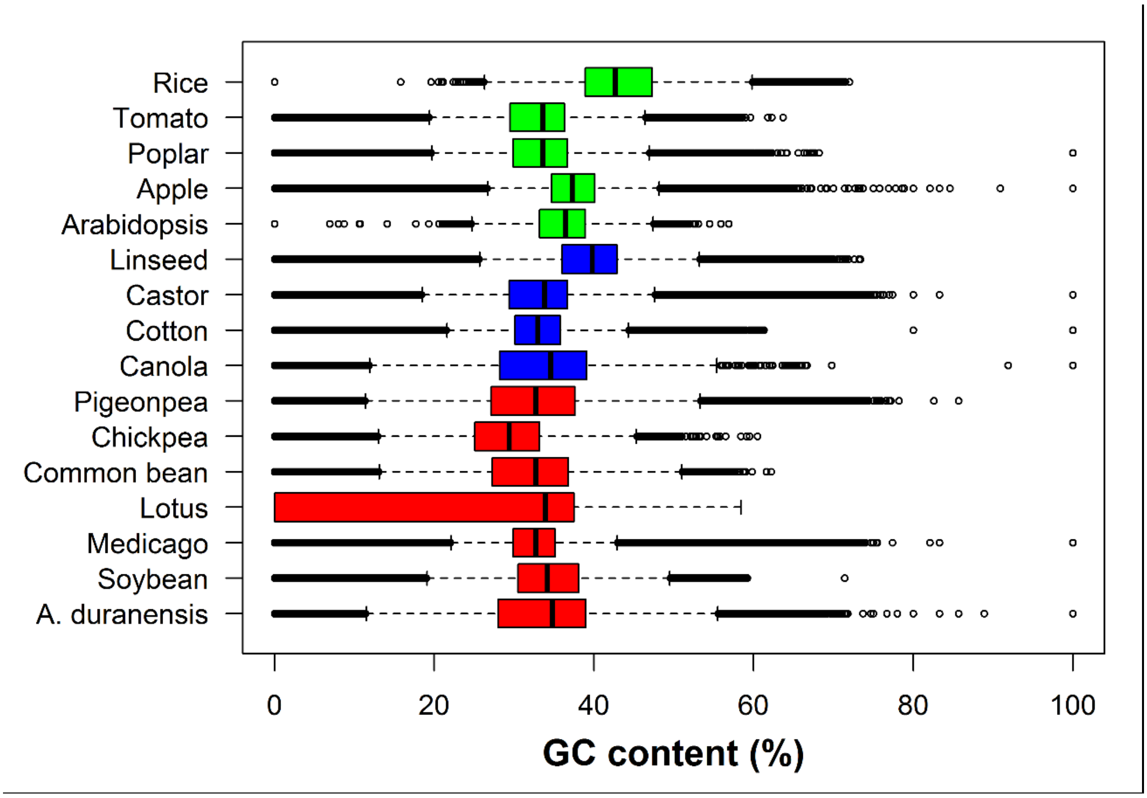


Figure S9. Comparison of the range of GC content among *A. duranensis* and other plant species.

The boxes display the likely range of the GC content variation (the interquartile range or IQR). The upper and lower bars represent upper and lower inner fences, respectively. The circles depict outliers in the GC content.

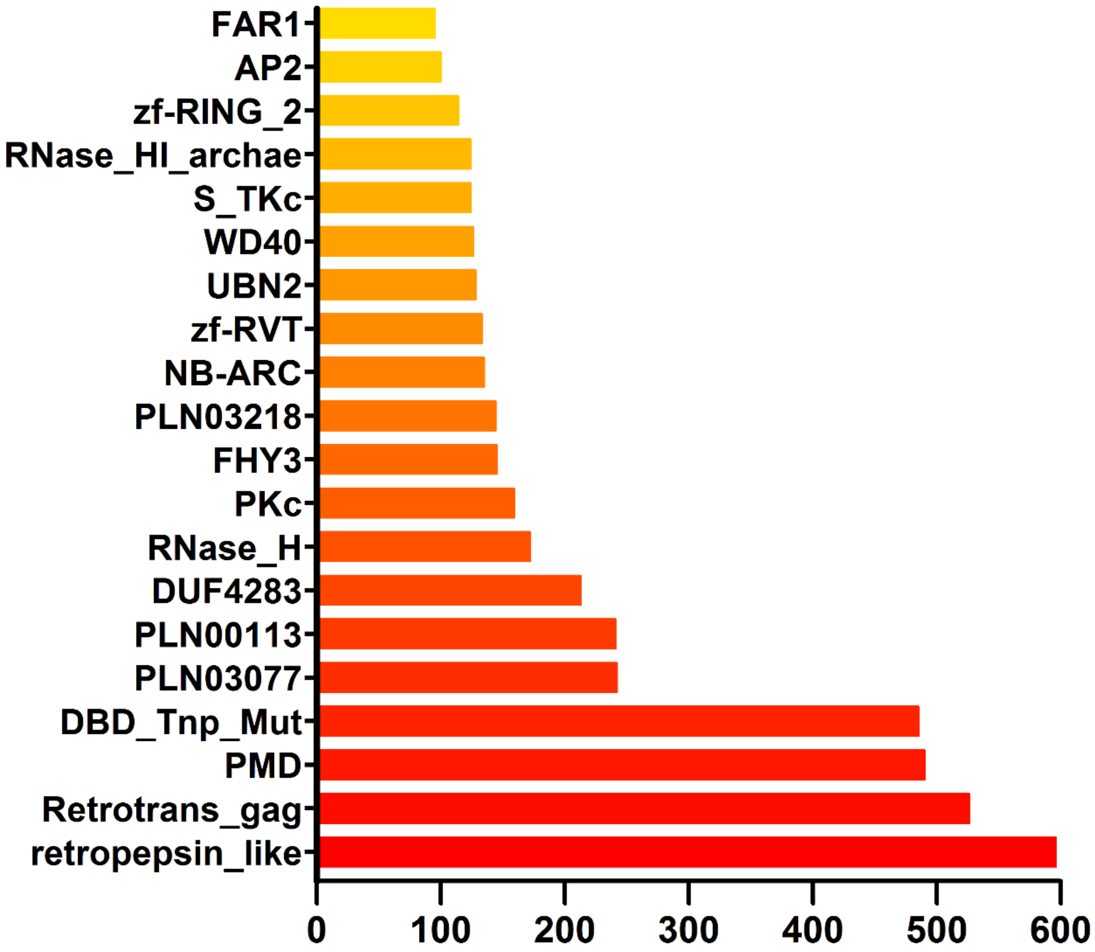


Figure S10. The top 20 Pfam domains for the *A. duranensis* genome.

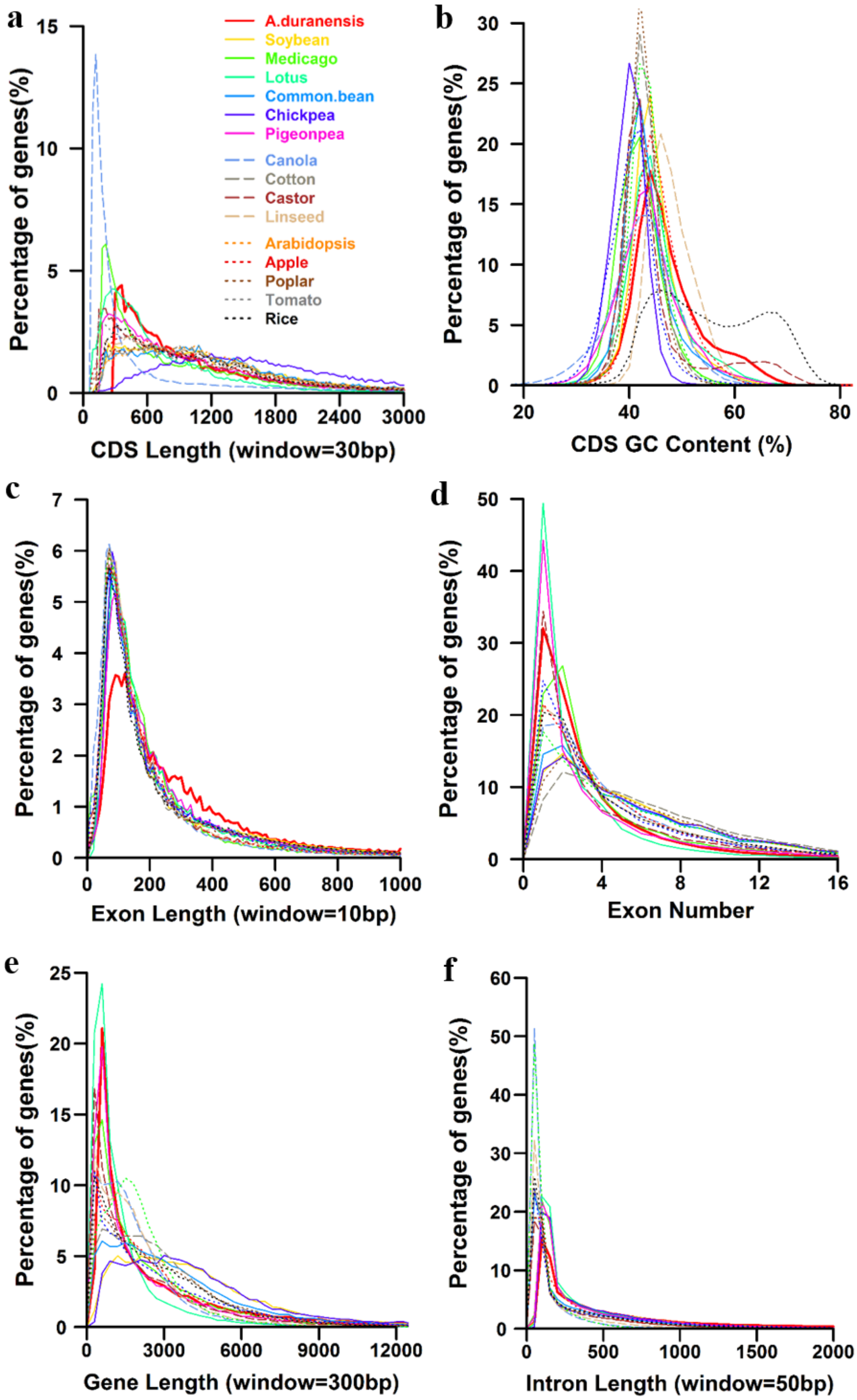


Figure S11. Distribution comparison of (a) CDS length, (b) CDs GC content, (c) exon length, (d) exon number, (e) gene length and (f) intron length among *A. duranensis* and other plant species.

The red solid line presents the distribution in *A. duranensis*. Solid lines represent legume species, dash lines represent oilseed crops and dot lines represent other plant species.

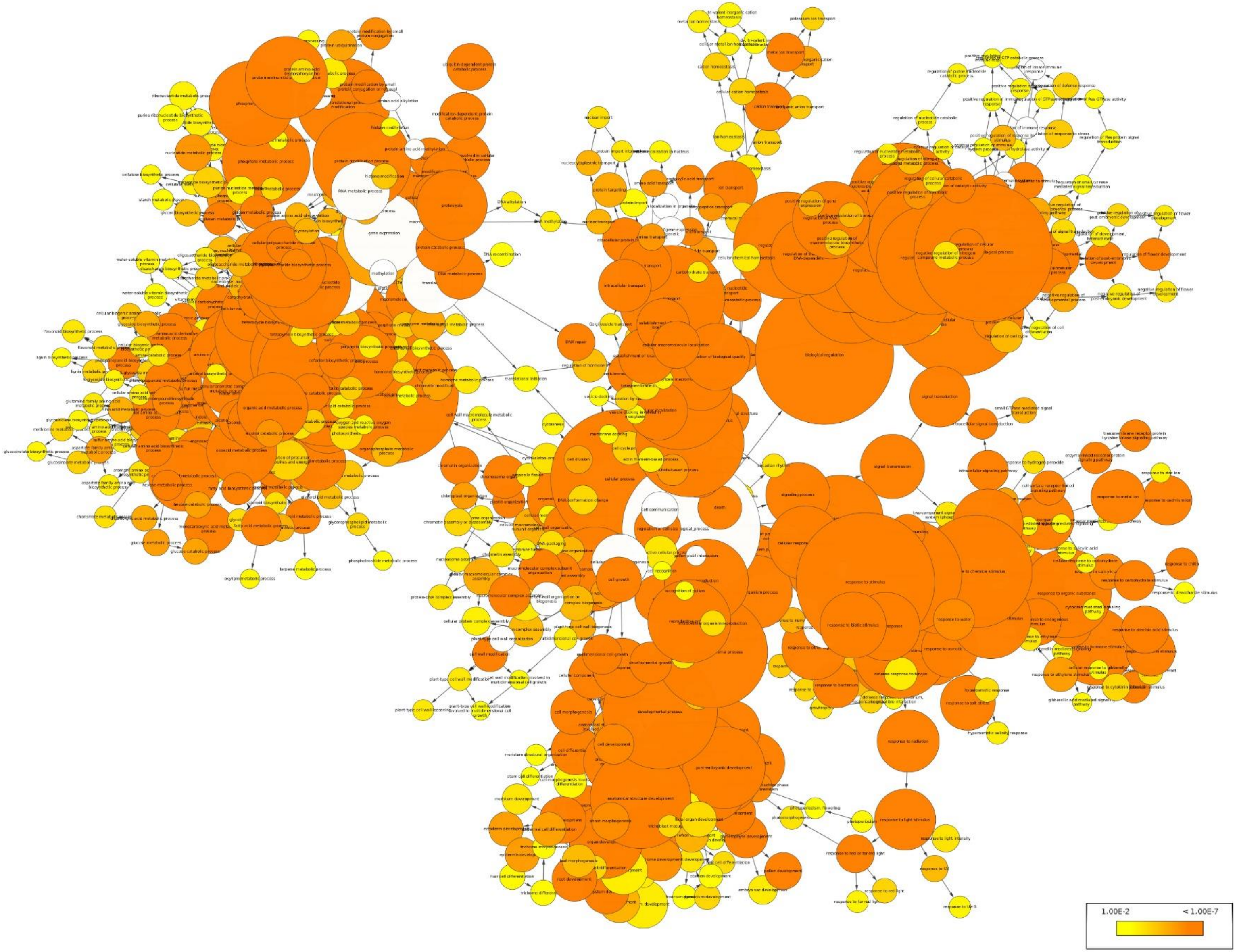


Figure S12. Enriched GO terms for biological process

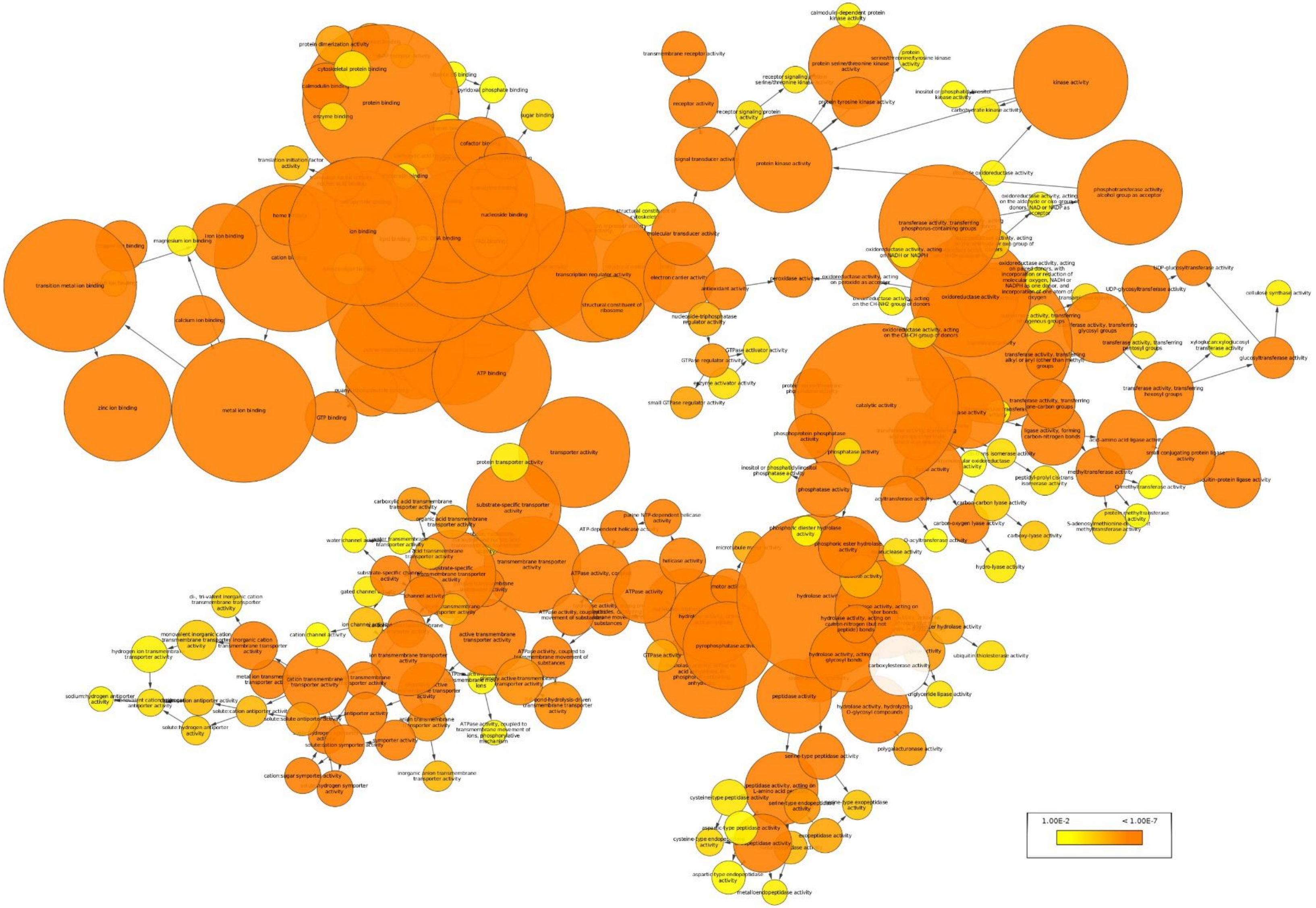

protein dimerization activity
cytoskeletal protein binding
calmodulin binding
protein binding
enzyme binding
pyridoxal phosphate binding
sugar binding
cofactor binding
translation initiation factor activity
heme binding
nucleoside binding
ion binding
magnesium ion binding
iron ion binding
cation binding
transition metal ion binding
calcium ion binding
zinc ion binding
metal ion binding
GTP binding
ATP binding
structural constituent of cytoskeleton
molecular transducer activity
transcription regulator activity
electron carrier activity
structural constituent of ribosome
transmembrane receptor activity
receptor activity
signal transducer activity
receptor signaling protein activity
receptor signaling protein serine/threonine kinase activity
calmodulin-dependent protein kinase activity
protein serine/threonine kinase activity
protein serine/threonine/tyrosine kinase activity
protein tyrosine kinase activity
protein kinase activity
kinase activity
inositol or phosphatidylinositol kinase activity
carbohydrate kinase activity
phosphotransferase activity, alcohol group as acceptor
oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
transferase activity, transferring phosphorus-containing groups
oxidoreductase activity, acting on NADH or NADPH
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NADH or NADPH as one donor, and incorporation of one atom of oxygen
oxidoreductase activity
antioxidant activity
peroxidase activity
oxidoreductase activity, acting on peroxide as acceptor
oxidoreductase activity, acting on the CH-NH2 group of donors
oxidoreductase activity, acting on the CH-CH group of donors
nucleoside-triphosphatase regulator activity
GTPase activator activity
GTPase regulator activity
enzyme activator activity
small GTPase regulator activity
UDP-glucosyltransferase activity
UDP-glycosyltransferase activity
cellulose synthase activity
transferase activity, transferring glycosyl groups
transferase activity, transferring pentosyl groups
xyloglucan:xyloglucosyl transferase activity
glucosyltransferase activity
transferase activity, transferring alkyl or aryl (other than methyl) groups
transferase activity, transferring hexosyl groups
catalytic activity
transferase activity, transferring one-carbon groups
ligase activity, forming carbon-nitrogen bonds
acid-amino acid ligase activity
small conjugating protein ligase activity
ubiquitin-protein ligase activity
phosphoprotein phosphatase activity
phosphatase activity
inositol or phosphatidylinositol phosphatase activity
peptidyl-prolyl cis-trans isomerase activity
methyltransferase activity
O-methyltransferase activity
protein methyltransferase activity
S-adenosylmethionine-dependent methyltransferase activity
acyltransferase activity
carbon-carbon lyase activity
carbon-oxygen lyase activity
carboxy-lyase activity
O-acyltransferase activity
hydro-lyase activity
transporter activity
protein transporter activity
substrate-specific transporter activity
carboxylic acid transmembrane transporter activity
organic acid transmembrane transporter activity
water channel activity
water transmembrane transporter activity
transmembrane transporter activity
substrate-specific channel activity
substrate-specific transmembrane transporter activity
gated channel activity
channel activity
di-, tri-valent inorganic cation transmembrane transporter activity
monovalent inorganic cation transmembrane transporter activity
inorganic cation transmembrane transporter activity
hydrogen ion transmembrane transporter activity
cation channel activity
ion channel activity
active transmembrane transporter activity
ion transmembrane transporter activity
cation transmembrane transporter activity
metal ion transmembrane transporter activity
sodium:hydrogen antiporter activity
solute:cation antiporter activity
solute:hydrogen antiporter activity
solute:solute antiporter activity
antiporter activity
anion transmembrane transporter activity
symporter activity
solute:cation symporter activity
cation:sugar symporter activity
inorganic anion transmembrane transporter activity
ATPase activity, coupled to transmembrane movement of substances
ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism
P-P-bond-hydrolysis-driven transmembrane transporter activity
purine NTP-dependent helicase activity
ATP-dependent helicase activity
helicase activity
ATPase activity
microtubule motor activity
motor activity
GTPase activity
pyrophosphatase activity
hydrolase activity
phosphoric diester hydrolase activity
phosphoric ester hydrolase activity
nuclease activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
hydrolase activity, acting on glycosyl bonds
carboxylesterase activity
ubiquitin thiolesterase activity
triglyceride lipase activity
hydrolase activity, hydrolyzing O-glycosyl compounds
polygalacturonase activity
peptidase activity
serine-type peptidase activity
peptidase activity, acting on L-amino acid peptides
cysteine-type peptidase activity
serine-type endopeptidase activity
serine-type exopeptidase activity
aspartic-type peptidase activity
exopeptidase activity
cysteine-type endopeptidase activity
endopeptidase activity
aspartic-type endopeptidase activity
metalloendopeptidase activity
1.00E-2
< 1.00E-7

Figure S13. Enriched GO terms for molecular functions

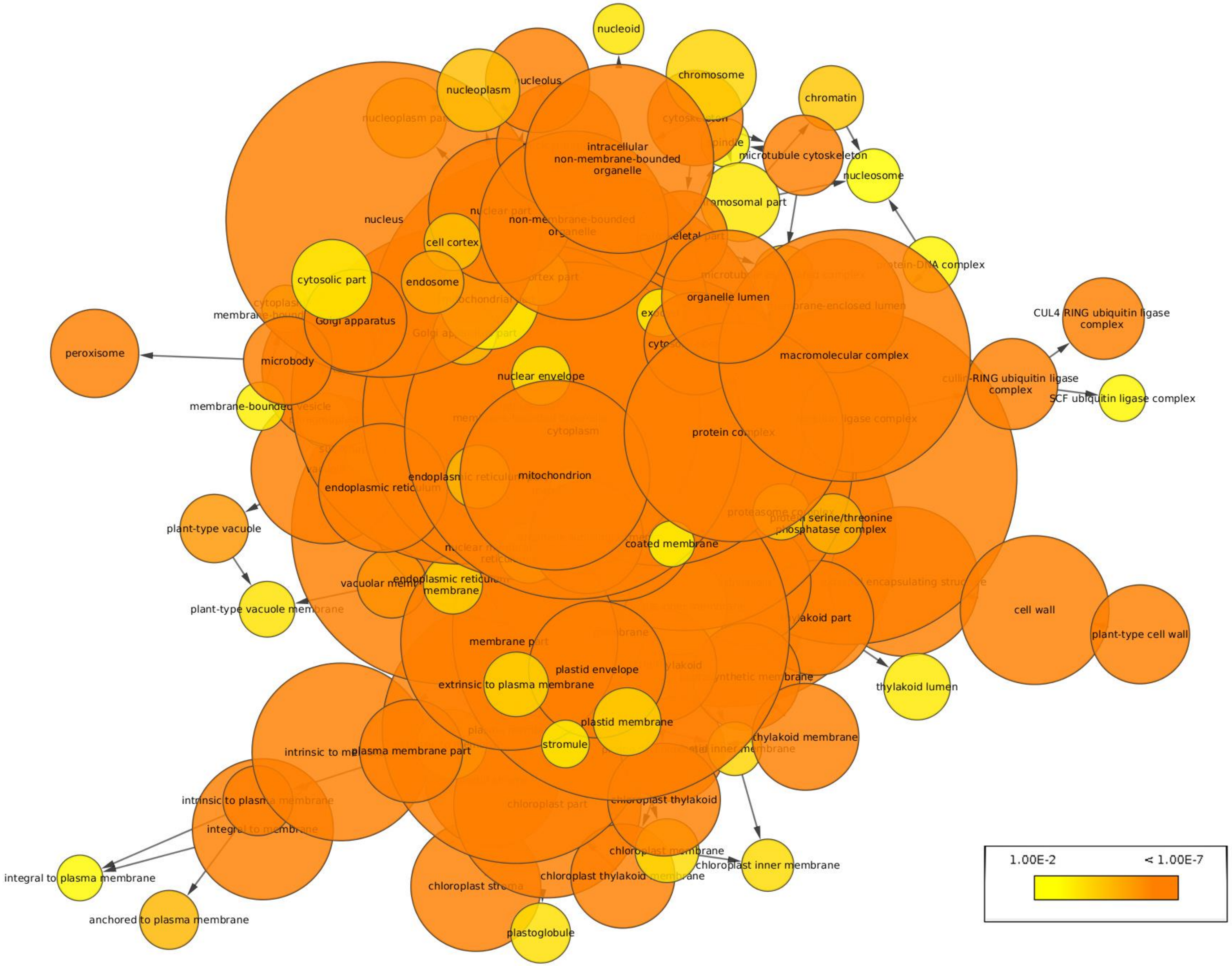


Figure S14. Enriched GO terms for cellular components

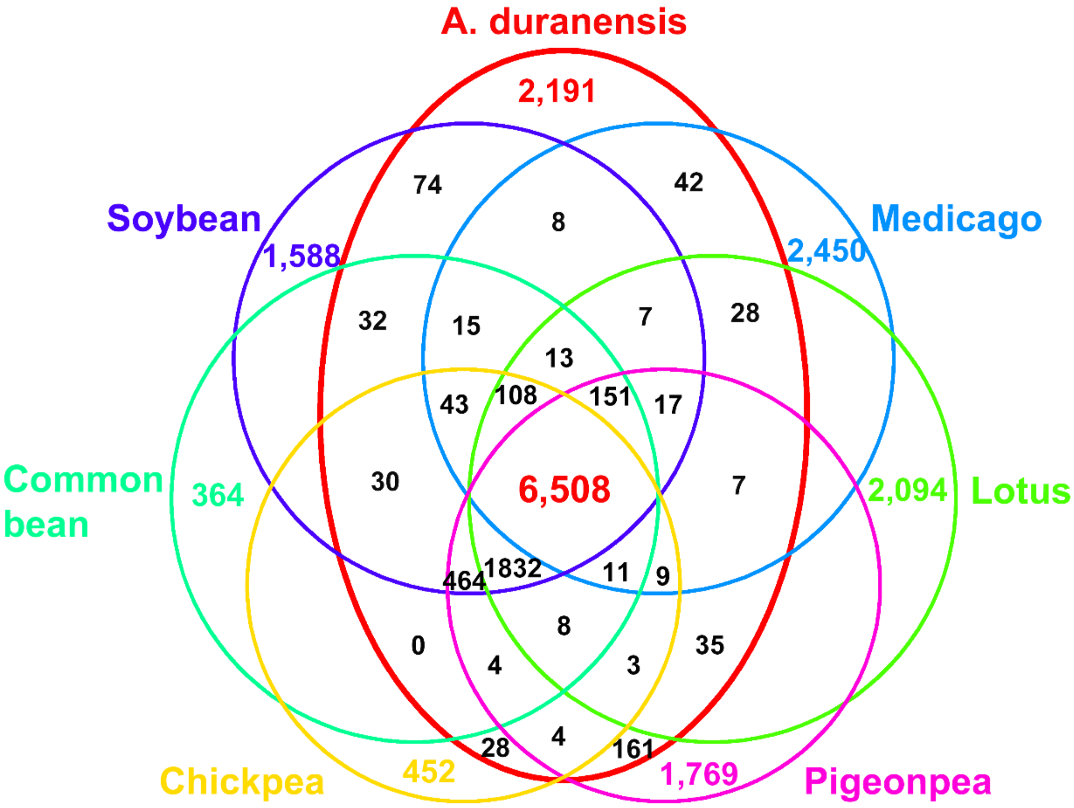


Figure S15. Venn diagram showing shared and unique gene families among legume crops.

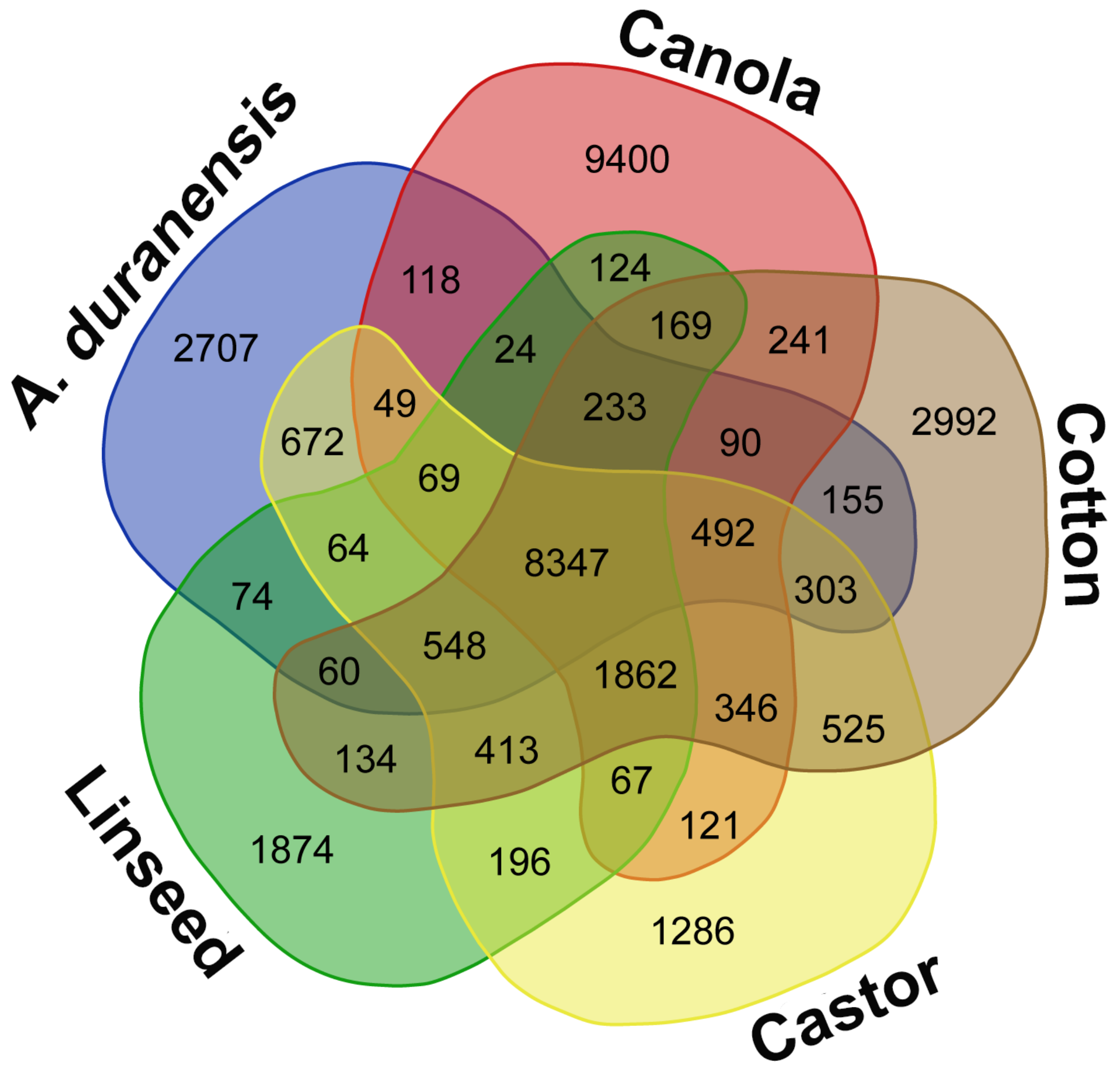


Figure S16. Venn diagram showing shared and unique gene families among oilseed crops.

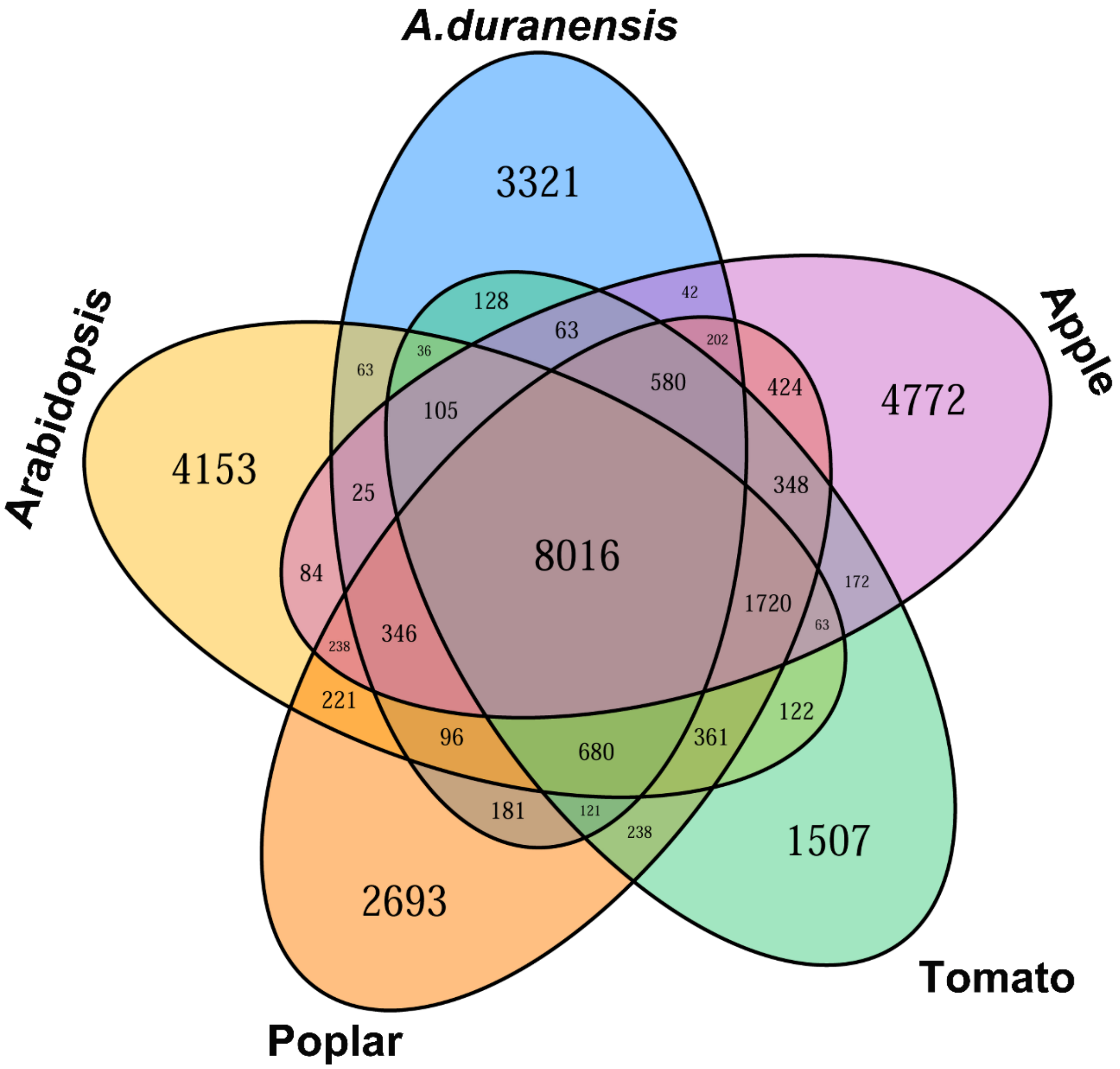


Figure S17. Venn diagram showing shared and unique gene families among distantly related plant species.

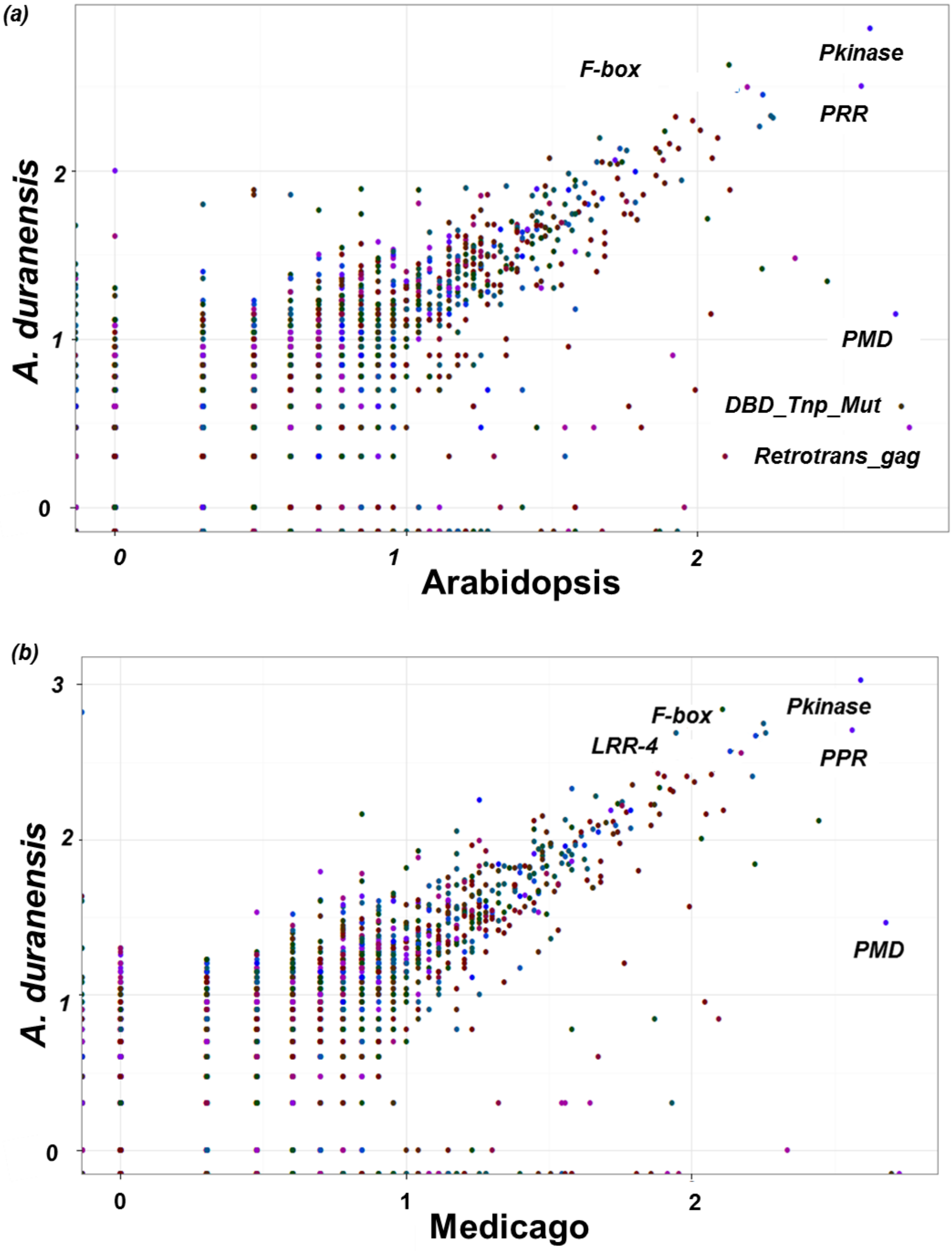


Figure S18. Pairwise scatterplot of gene family members between *A. duranensis* and *Arabidopsis* as well as *Medicago*.

The number of members in each family are log10 transformed, and then plotted pairwise. The values >2.5 are only labelled to ease visualization.

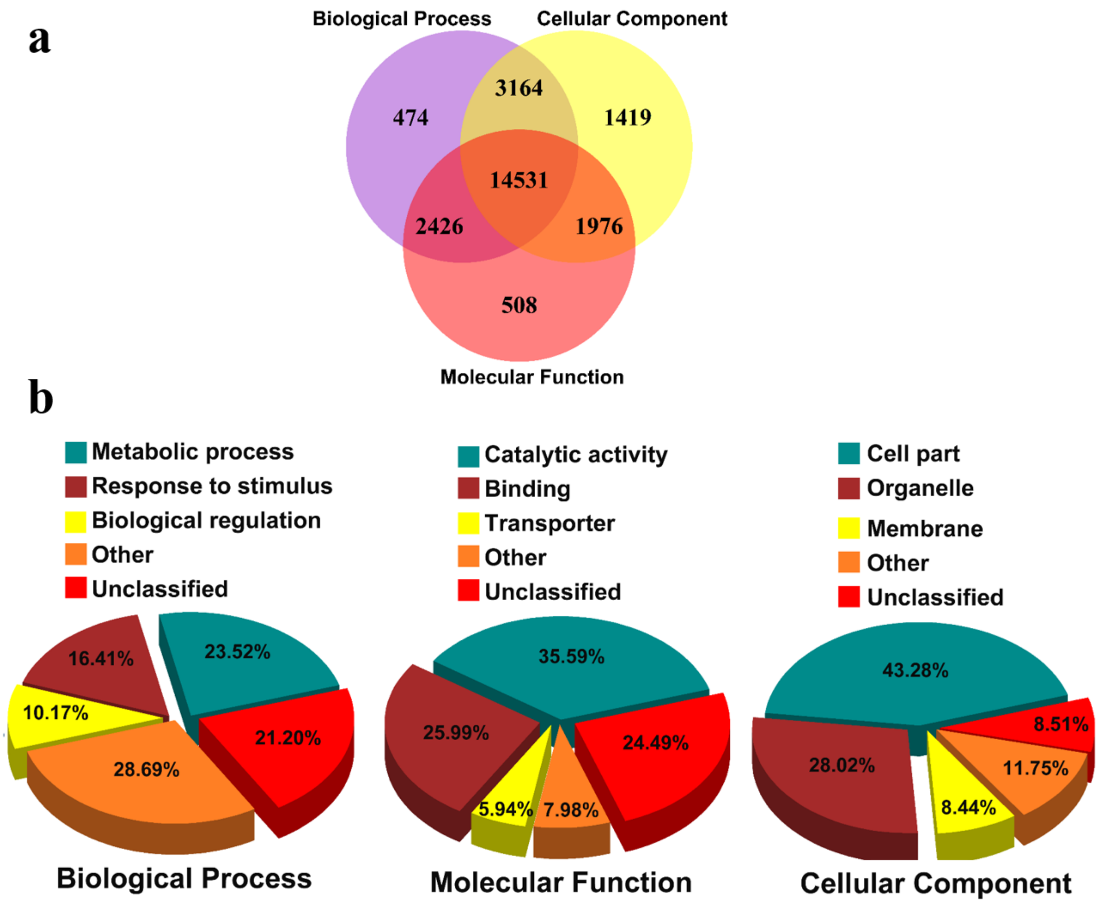


Figure S19. Venn diagram of GO annotation (overlapping genes among three ontologies) in *A. duranensis* predicted protein-coding genes.

This figure shows (a) the intersection and relationship of each ontology, and (b) the fractions for the top 4 categories in each ontology and the remaining categories.

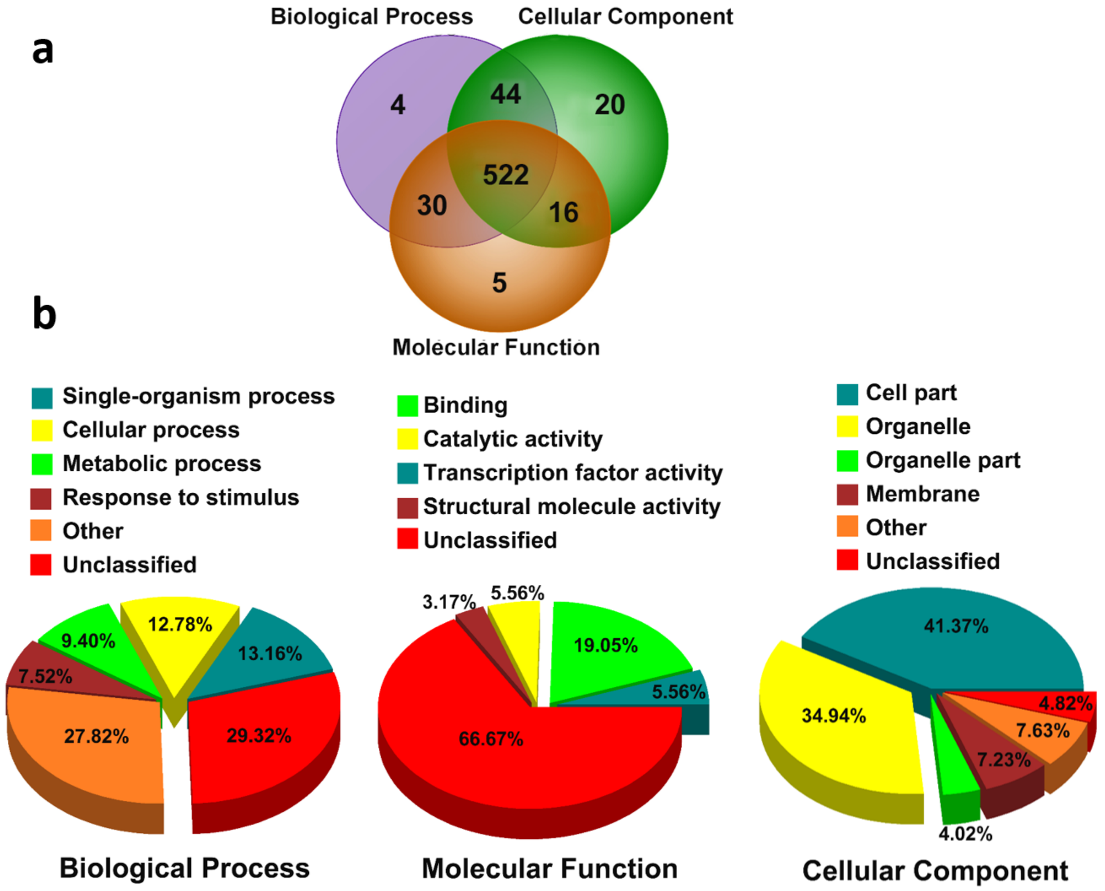


Figure S20. Venn diagram of GO annotation (overlapping genes among three ontologies) in *A. duranensis* specific genes.

This figure shows (a) the intersection and relationship of each ontology, and (b) the fractions for the top 4 or 5 categories in each ontology and the remaining categories.

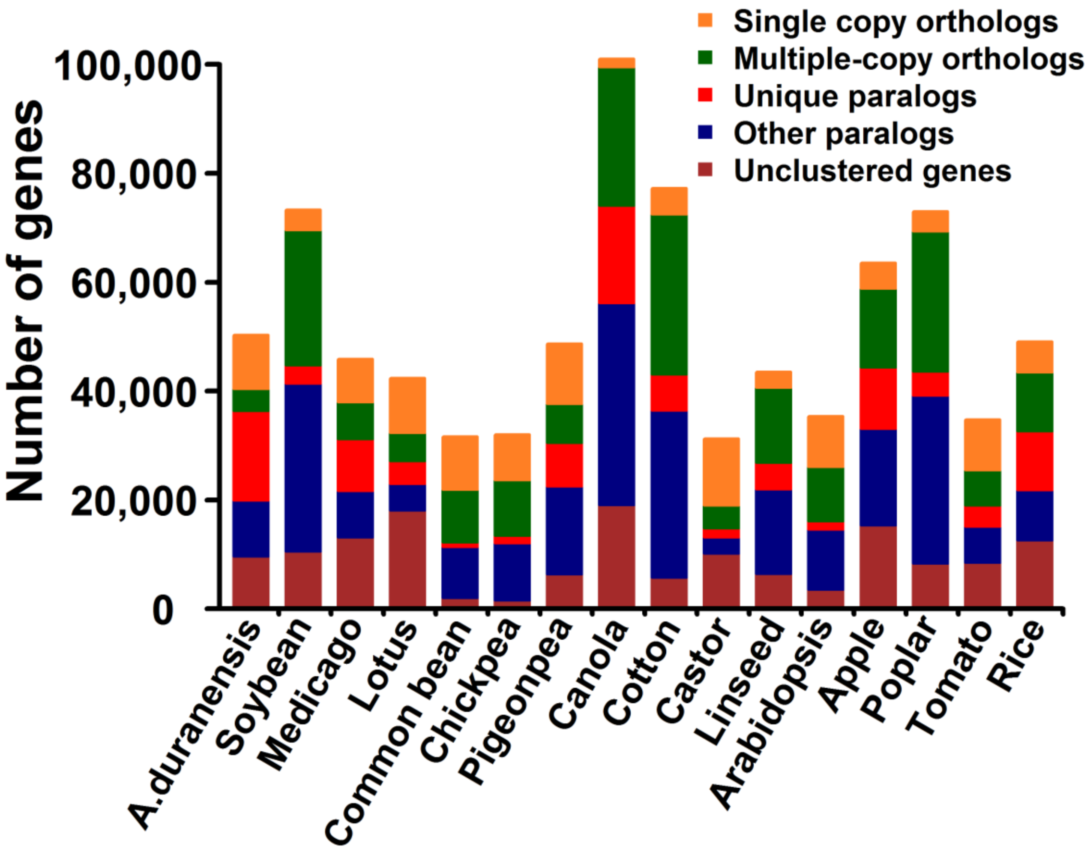


Figure S21. Comparison of orthologous genes among *A. duranensis* and other plant species.

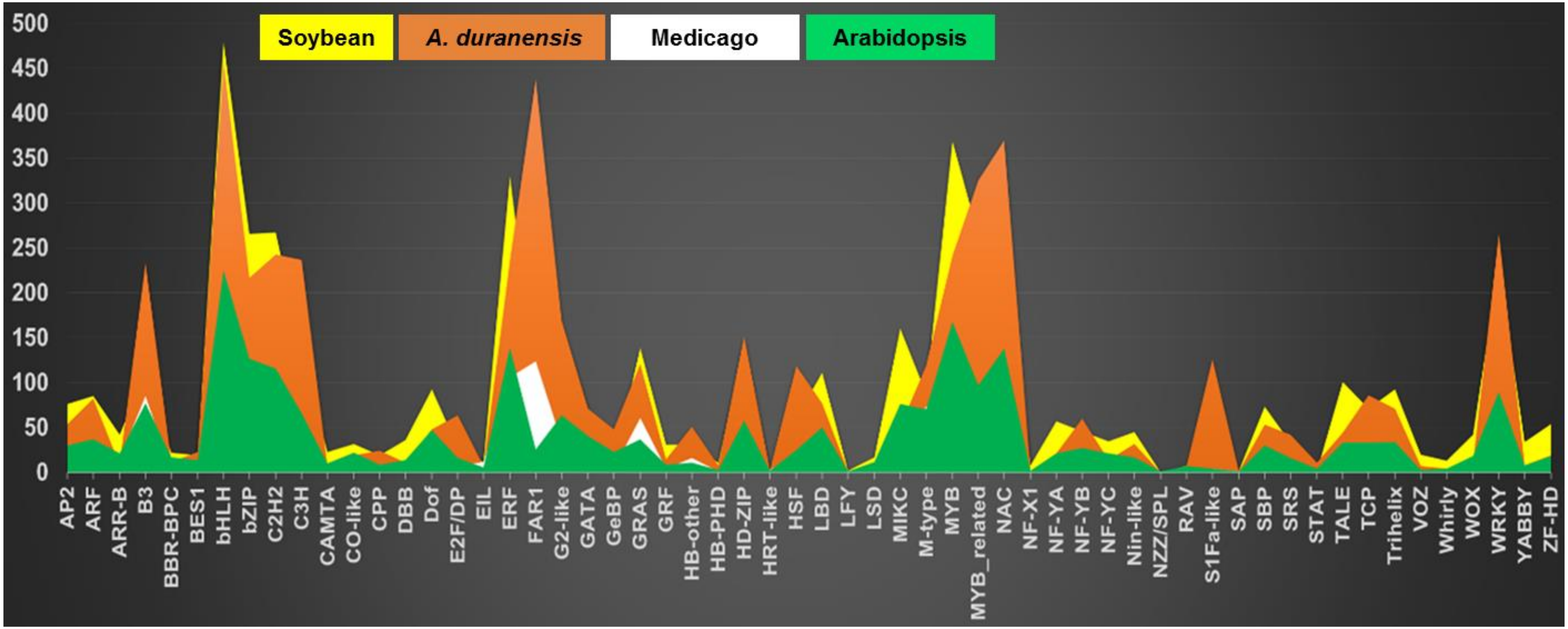


Figure S22. Distribution of TF genes in different TF families among the four species.

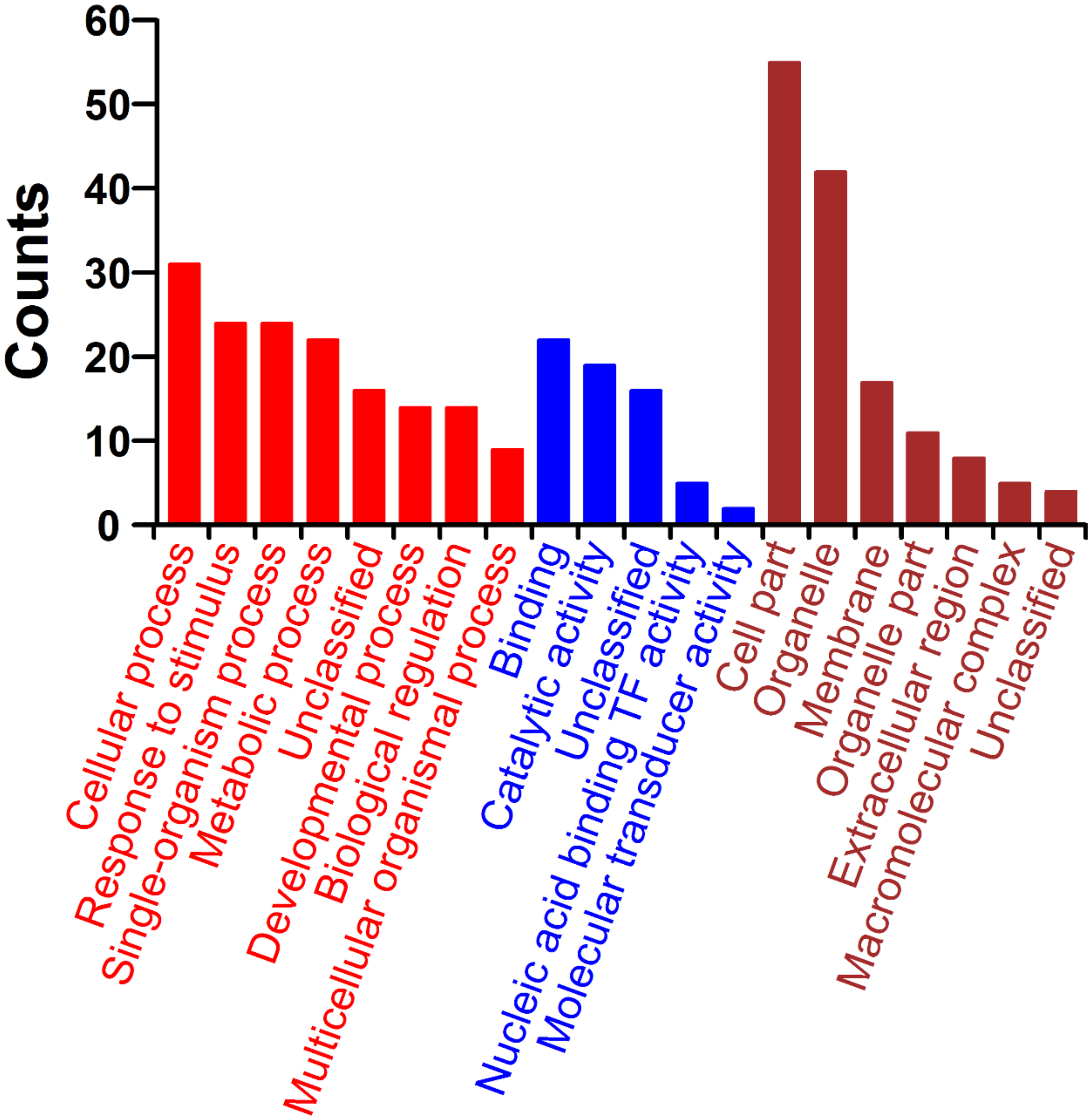


Figure S23. GO classification of miRNA target genes in *A. duranensis*.

Red colors represent categories of Cellular Component, blue colors represent categories of Biological Process, and brown colors represent categories of Molecular Function.

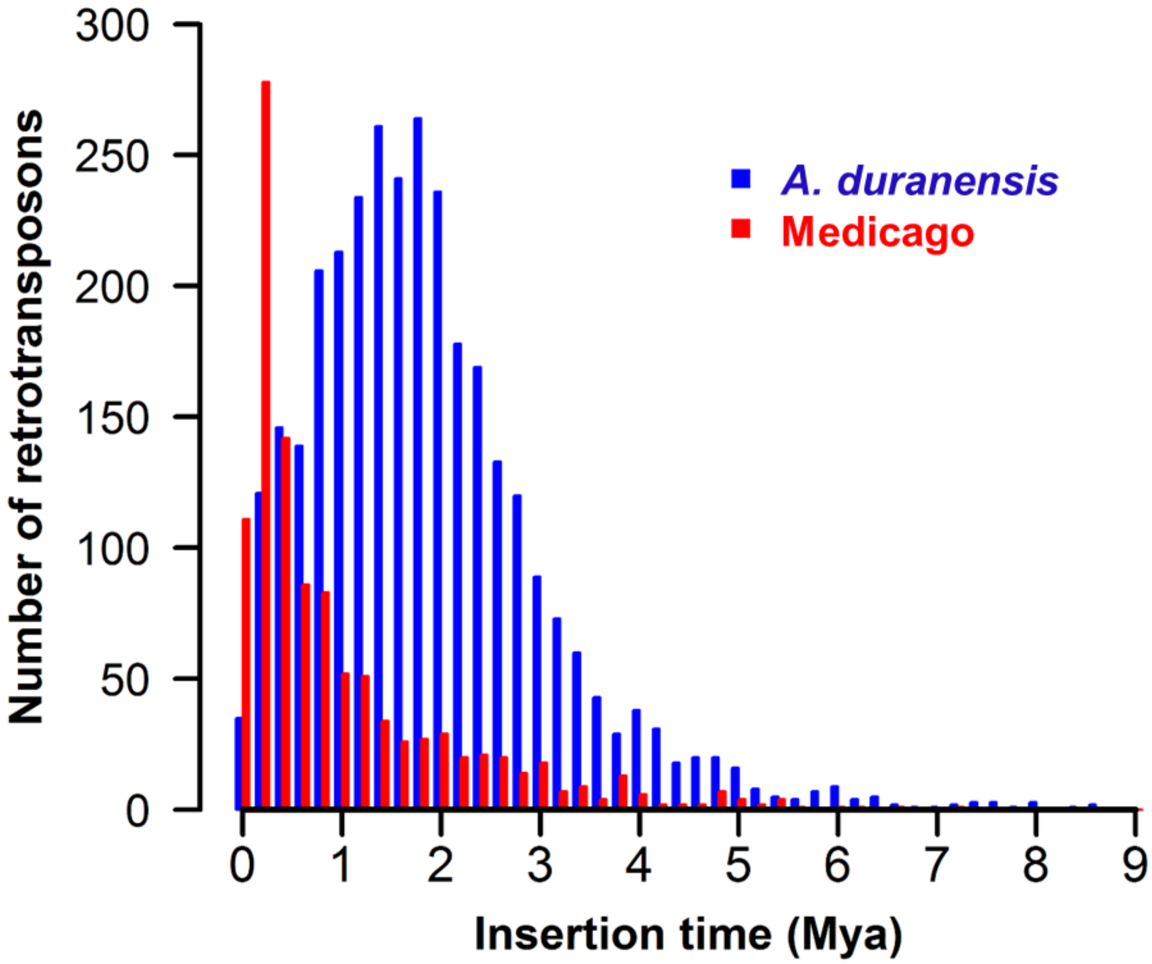


Figure S24. Dating the LTR retrotransposon insertion time. Dating of *M. truncatula* LTR retrotransposons was used as a comparison.

LTR retrotransposon sequences found by LTR_finder were clustered by CD-HIT at 90% of sequence similarity with 90% coverage of the shorter sequence. The LTR sequences were not included in the calculation of sequence similarity and coverage. The longest in each cluster was selected as the representative member, of which LTRs were aligned and transitions and transversions were computed and used for the insertion time computation.

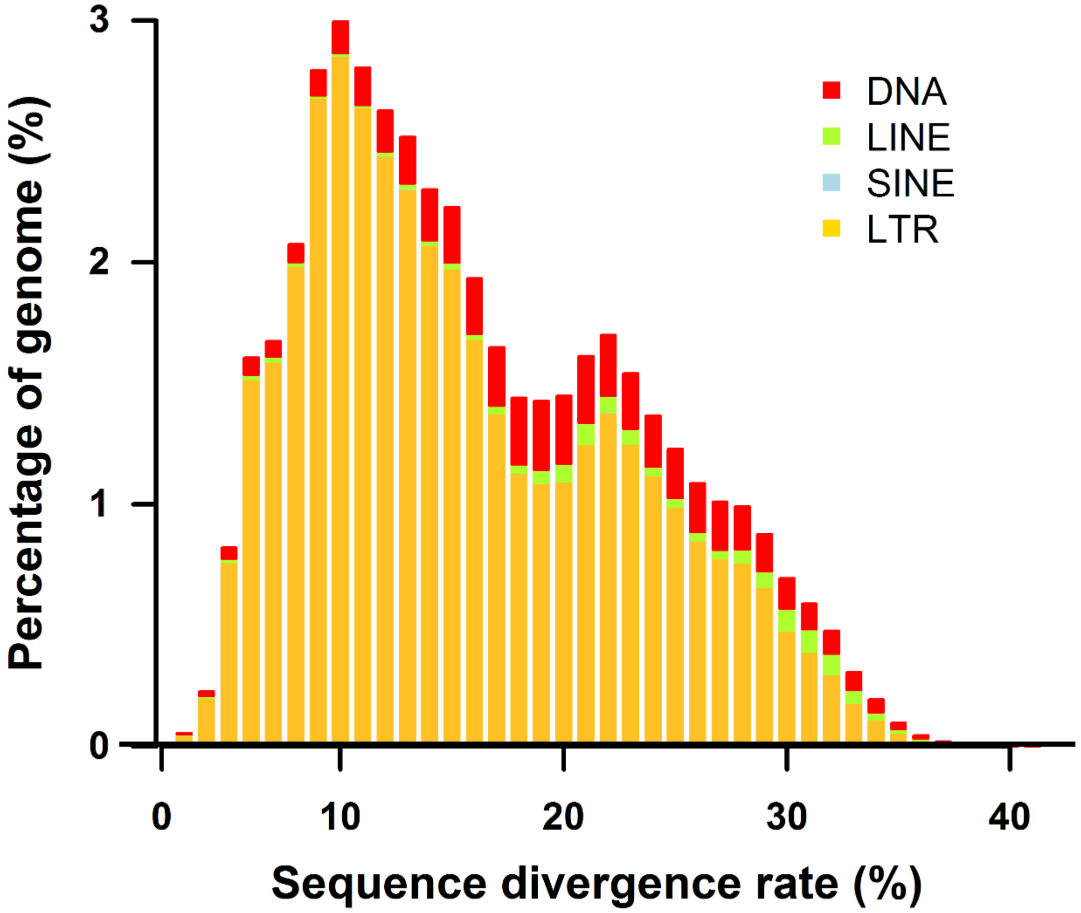


Figure S25. Distribution of divergence rate of different repetitive elements (DNA elements, LINE, LTR, SINE) in the *A. duranensis* genome.

Divergence rate was calculated between the identified TEs in the genome and the consensus sequence in the TE library (Repbase: http://www.girinst.org/repbase). DNA, DNA elements; LINE, long interspersed nuclear elements; LTR, long terminal repeat transposable element; SINE, short interspersed nuclear elements.

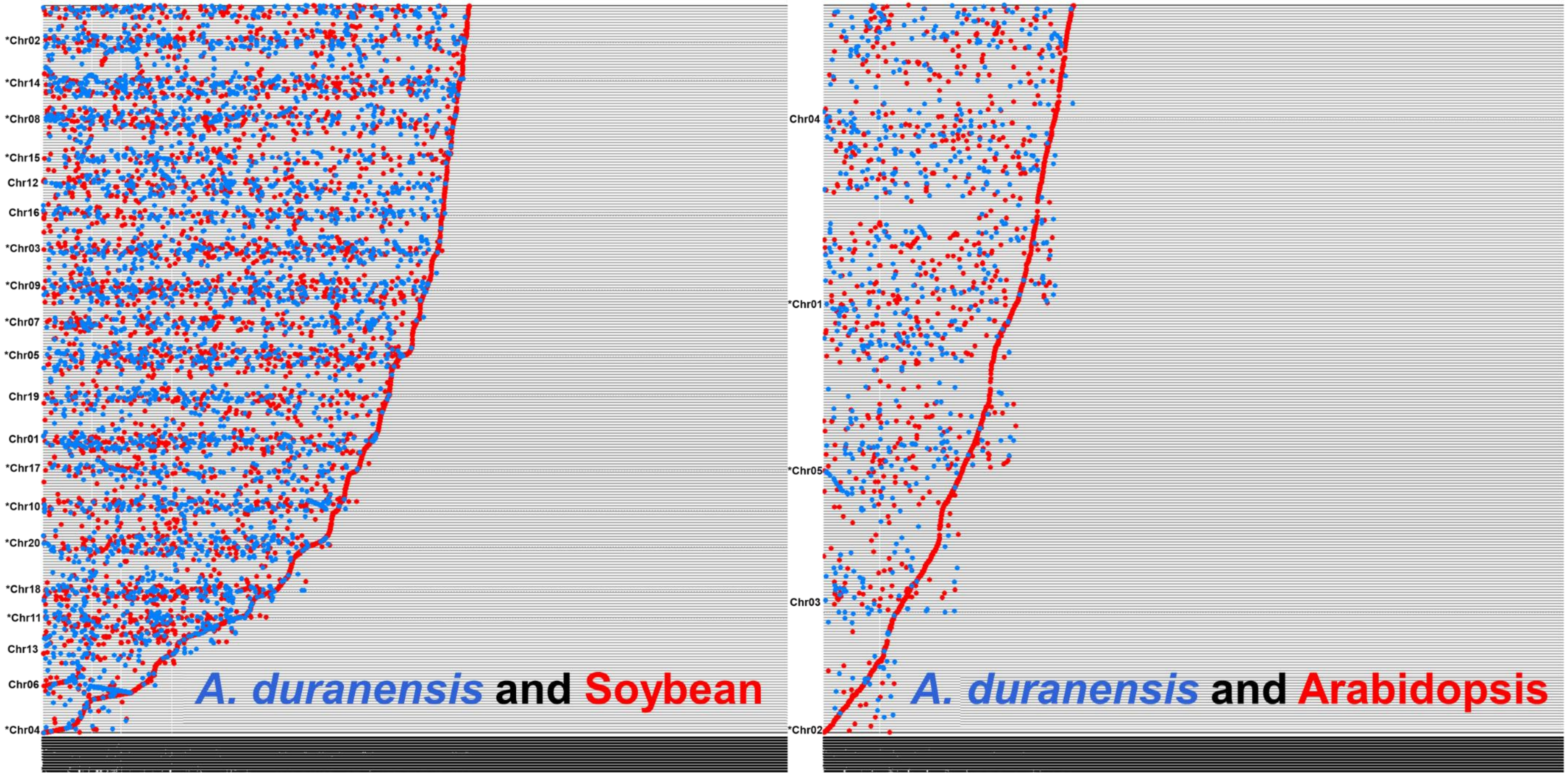


Figure S26. Syntenic blocks between *A. duranensis* scaffolds and Soybean and Arabidopsis chromosomes.

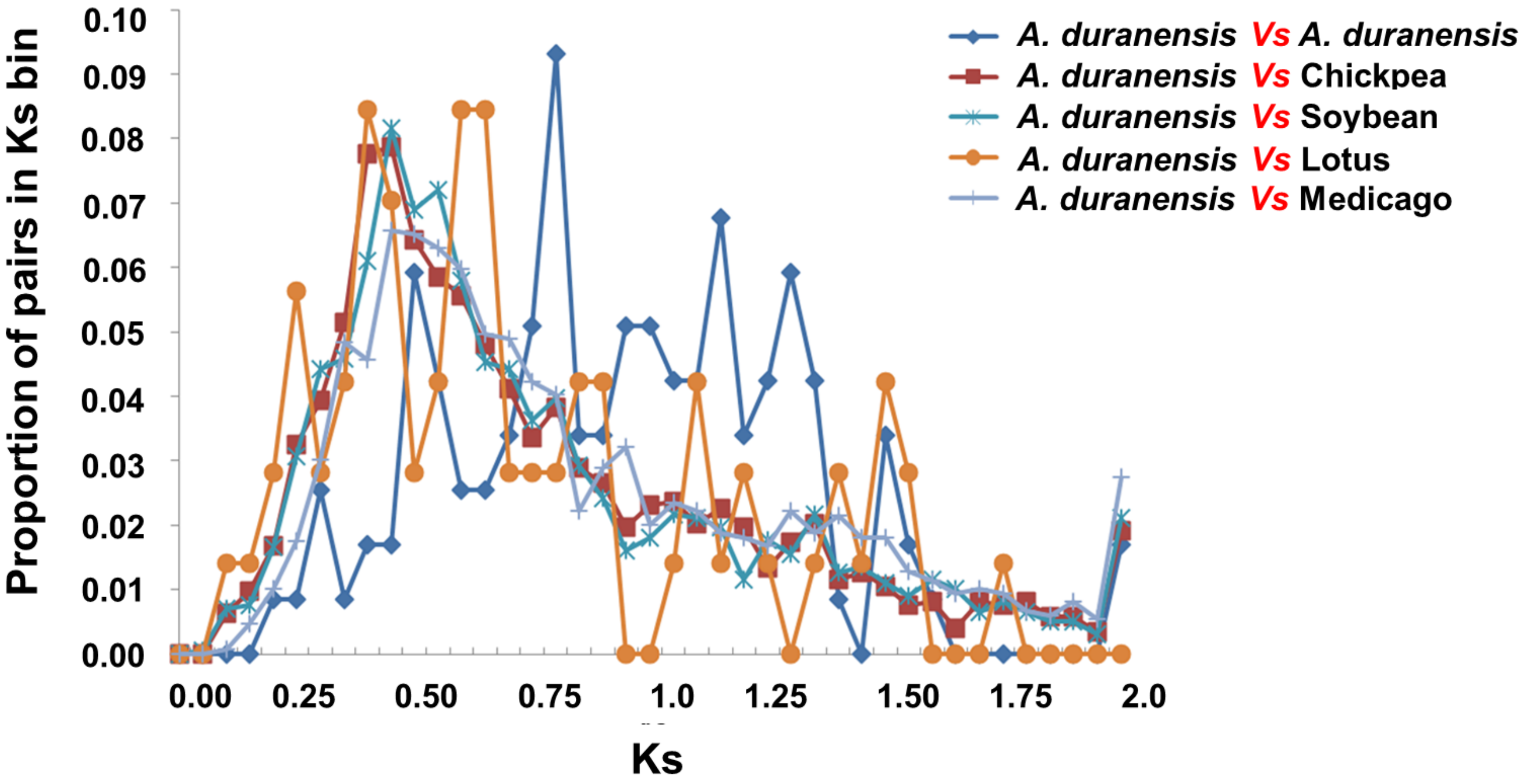


Figure S27. ***Ks*** **analysis of legume species.**

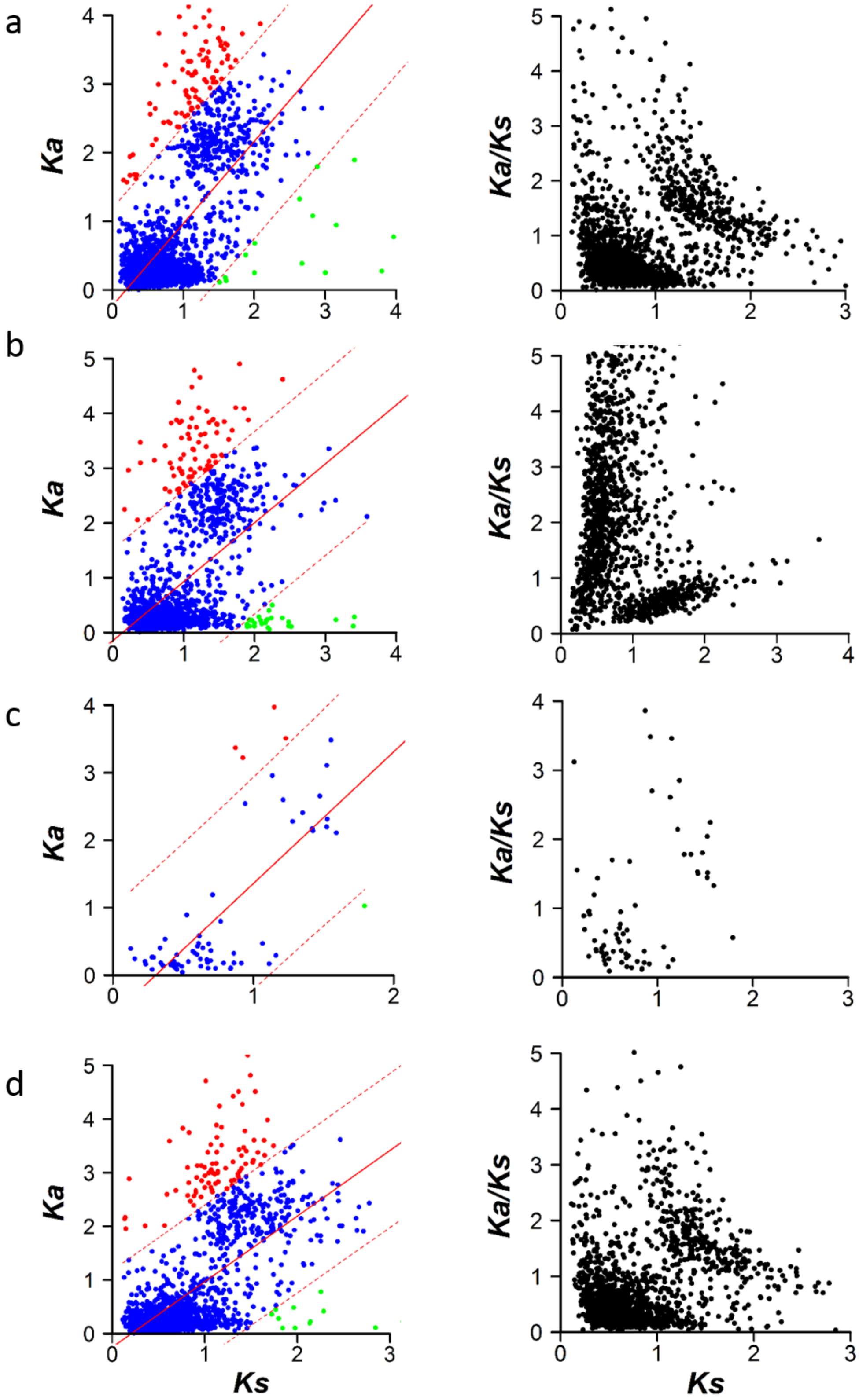


Figure S28. Scatterplot of *Ks* vs. *Ka* of orthologs between *A. duranensis* and soybean (a), *Medicago* (b), *Lotus* (c) and pigeonpea (d).

The dashed line represents the prediction interval about the linear regression. Red and green dots represent high and low ω (Ka/Ks) gene pairs, respectively.

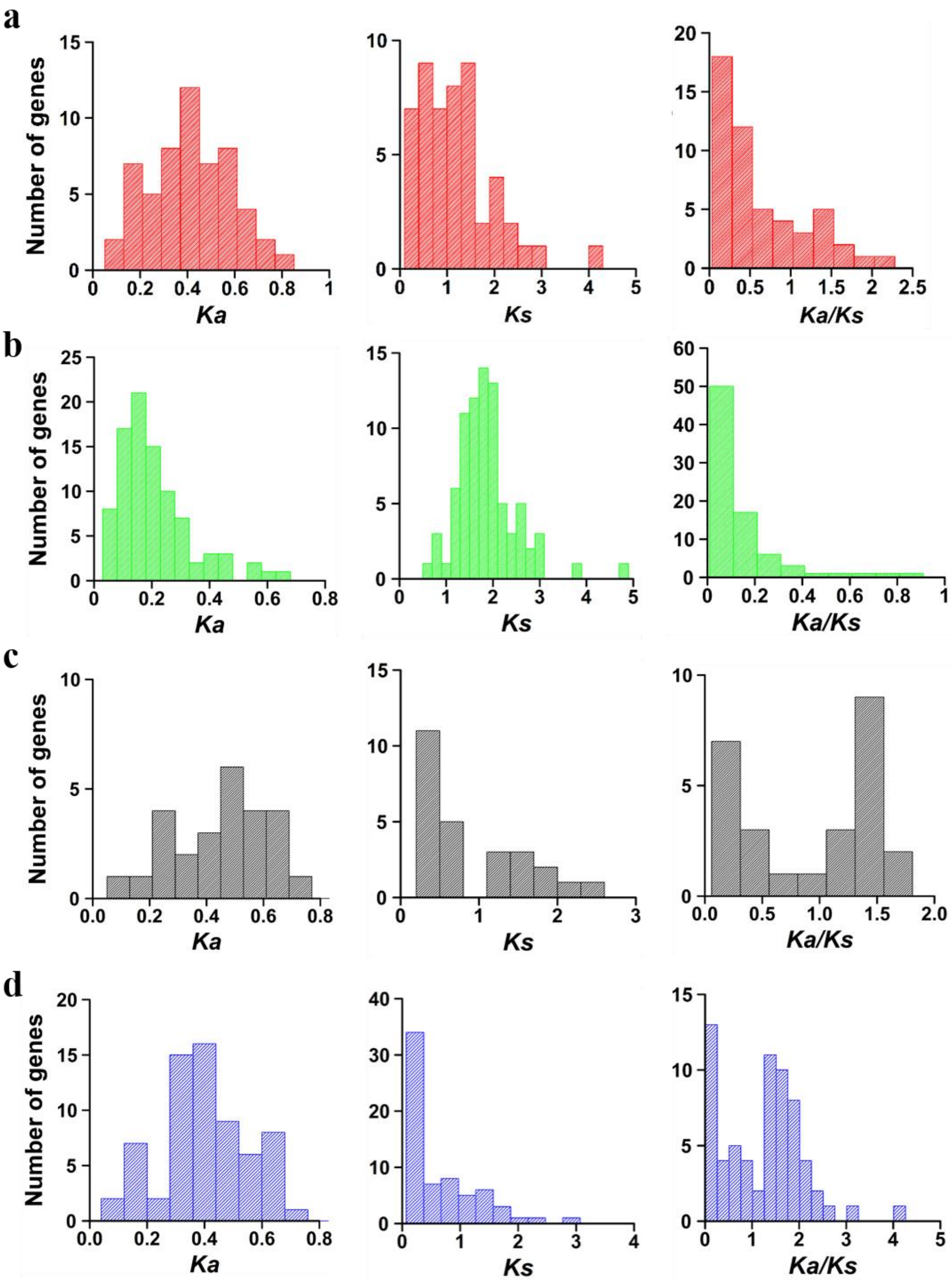
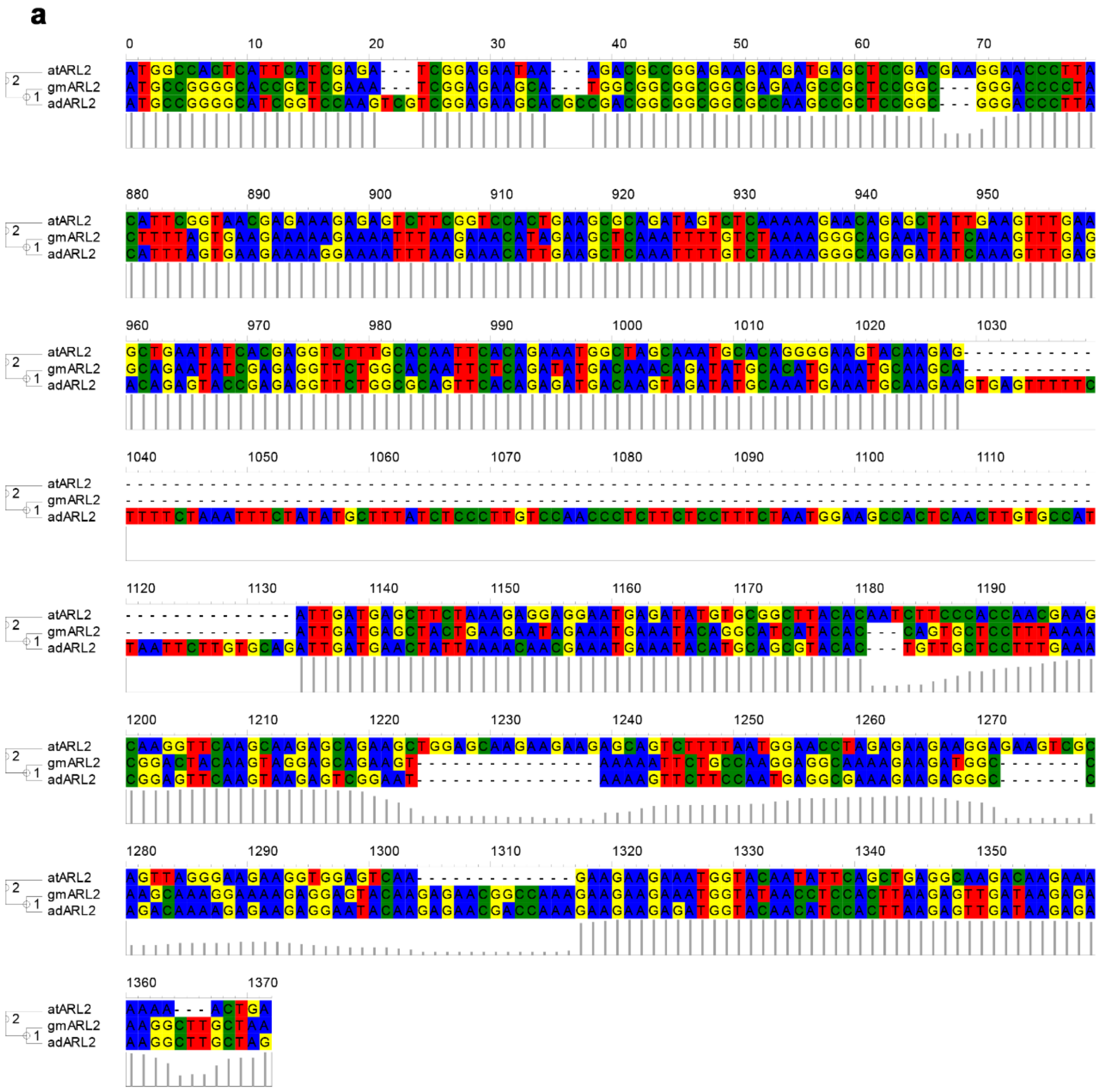


Figure S29. Distribution of *Ka, Ks* and ω (*Ka/Ks*) in pairs of (a) *Arabidopsis* and *A. duranensis* genes involved in gravitropism as well as in (b) *Arabidopsis* and

soybean. (c) Distribution of *Ka*, *Ks* and ω in pairs of *Arabidopsis* and *A. duranensis* (c) genes related to photomorphogenesis as well as of (d) *Arabidopsis* and soybean.


a
0 10 20 30 40 50 60 70
880 890 900 910 920 930 940 950
960 970 980 990 1000 1010 1020 1030
1040 1050 1060 1070 1080 1090 1100 1110
1120 1130 1140 1150 1160 1170 1180 1190
1200 1210 1220 1230 1240 1250 1260 1270
1280 1290 1300 1310 1320 1330 1340 1350
1360 1370
atARL2
gmARL2
adARL2
2
1

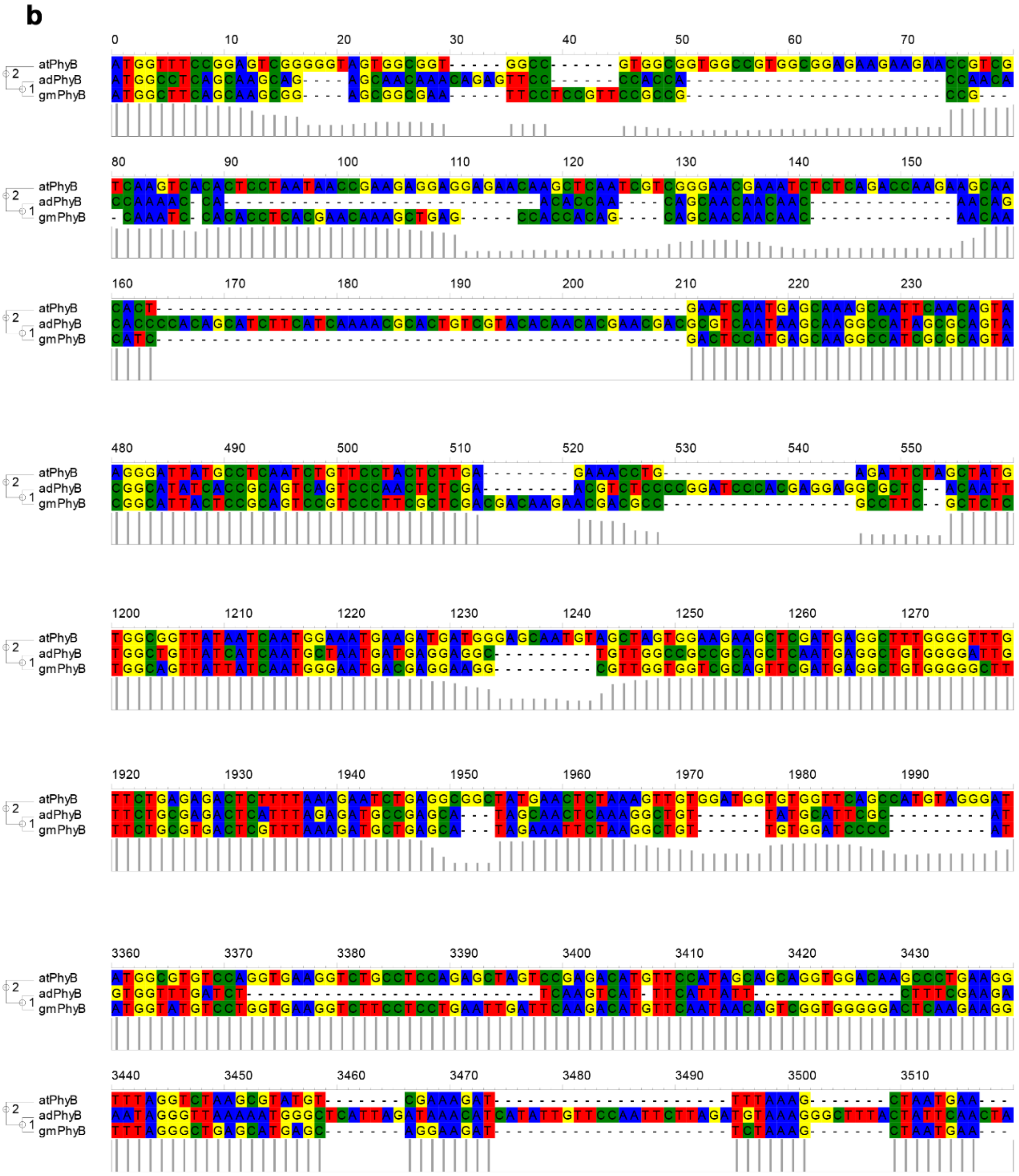


Figure S30. SNPs and Indels of representative genes under positive selection for gravitropism *ARL2* (a) and photomorphogenesis *phyB* (b) in *A. duranensis* (ad), *Arabidopsis* (at) and soybean (gm). Phylogeny-aware alignments of these genes were performed using PRANK and visualized using PRANKSTER. The approximate guide trees is indicated in left for each alignment. Alignments resulting in large-effect indel are shown.

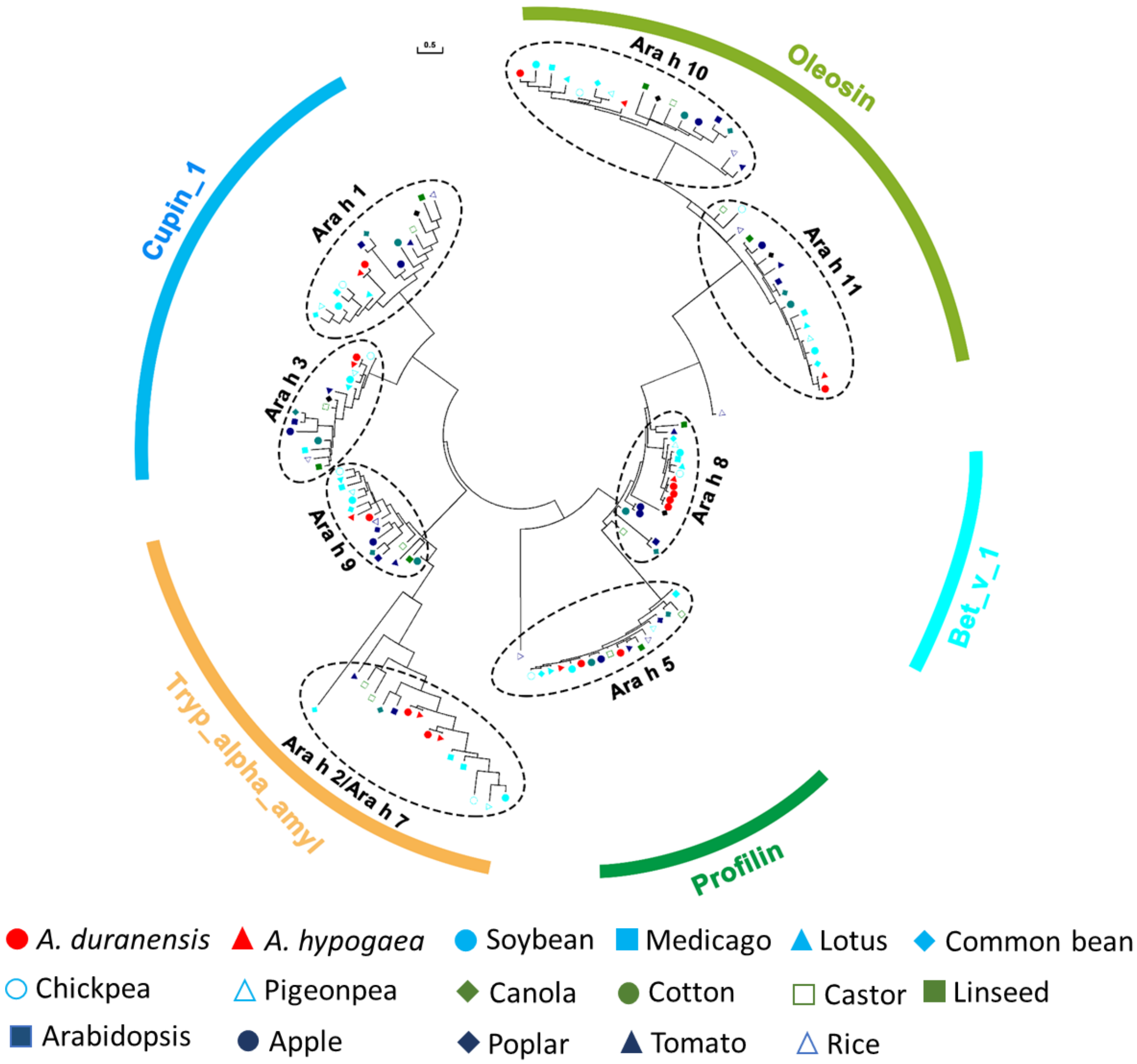


Figure S31. Phylogenetic tree of *Ara h 1-11* allergens, including sequences from previously identified homologs from cultivated peanut and other species.

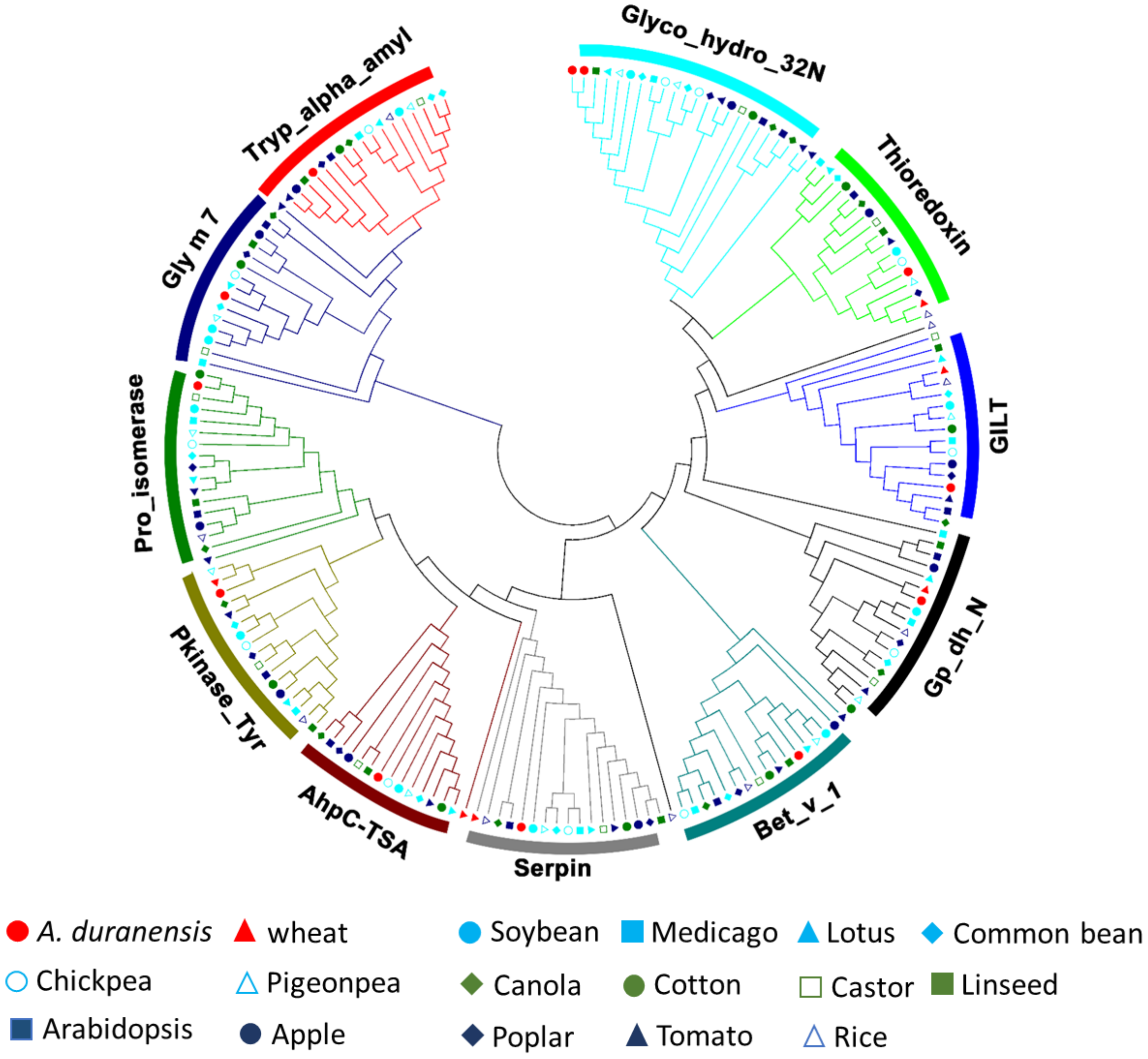


Figure S32. Phylogenetic tree of newly identified putative allergens in *A. duranensis* and the homologous proteins in other plant species.

SI References

1. Stalker, H.T. et al. Genetic diversity within the species *Arachis duranensis* Krapov. & W.C. Gregory, a possible progenitor of cultivated peanut. *Genome* **38**, 1201-1212 (1995).
2. Simpson, C.E. et al. History of *Arachis* including evidence of *A. hypogaea* L. progenitors. *Peanut Sci* **28**, 78-80 (2001).
3. Doyle, J.J. & Doyle, J.L. Isolation of plant DNA from fresh tissue. *Focus (Gico-BRL)* **12**, 13-15 (1990).
4. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18, doi:10.1186/2047-217X-1-18 (2012).
5. Boetzer, M. et al. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-9 (2010).
6. Li, R. et al. The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-7 (2010).
7. Chen, X. et al. Deep sequencing analysis of the transcriptomes of peanut aerial and subterranean young pods identifies candidate genes related to early embryo abortion. *Plant Biotechnol J* **11**, 115-127 (2013).
8. Ning, Z. et al. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-1729 (2001).
9. Temsch, E.M. & Greilhuber, J. Genome size in *Arachis duranensis*: a critical study. *Genome* **44**, 826-30 (2001).
10. Cantarel, B.L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188-96 (2008).
11. Marchler-Bauer, A. et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**, D222-6 (2015).
12. Finn, R.D. et al. Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-230 (2014).
13. Sonnhammer, E.L. et al. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-20 (1997).
14. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-61 (2004).
15. Kanehisa, M. et al. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-80 (2004).
16. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-40 (2014).
17. Zdobnov, E.M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-8 (2001).
18. Maere, S. et al. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-9 (2005).
19. Purugganan, M.D. et al. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. *Genetics* **140**, 345-56 (1995).
20. Li, L. et al. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-89 (2003).
21. Lavin, M. et al. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst Biol* **54**, 575-94 (2005).
22. Burge, S.W. et al. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* **41**, D226-232 (2012).
23. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).
24. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100-8 (2007).
25. Nawrocki, E.P. et al. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335-7 (2009).
26. Jeong, D.H. et al. Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. *Plant Cell* **23**, 4185-207 (2011).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
28. Smit, A.F.A. & Hubley, R. RepeatModeler Open-1.0. (2010). http://www.repeatmasker.org
29. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467 (2005).
30. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-82 (2007).
31. Piegu, B. et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res* **16**, 1262-9 (2006).
32. SanMiguel, P. et al. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20**, 43-5 (1998).
33. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
34. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-20 (1980).

35. Ma, J. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* **101**, 12404-10 (2004).
36. Thiel T, Michalek W, Varshney RK, Graner A: Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor Appl Genet* **106(3)**, 411-422 (2003).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
38. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
39. Kumar, S. et al. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* **9**, 299-306 (2008).
40. Young, N.D. et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520-4 (2011).
41. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183 (2010).
42. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-7 (2007).
43. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, e49 (2012).
44. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-91 (2007).
45. Guo, H. et al. Extensive and biased intergenomic nonreciprocal DNA exchanges shaped a nascent polyploid genome, *Gossypium* (cotton). *Genetics* **197**, 1153-63 (2014).
46. Kurtz, S. et al. Versatile and open software for comparing large genomes, *Genome Biology*, **5**, R12 (2004).
47. Beckstette, M. et al. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* **7**, 389 (2006)
48. Proost, S. et al. i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acid Research* **40(2)**, e11 (2012).
49. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).
50. Kato, T. et al. SGR2, a phospholipase-like protein, and ZIG/SGR4, a SNARE, are involved in the shoot gravitropism of *Arabidopsis*. *Plant Cell* **14**, 33-46 (2002).
51. Yano, D. et al. A SNARE complex containing SGR3/AtVAM3 and ZIG/VTI11 in gravity-sensing cells is important for *Arabidopsis* shoot gravitropism. *Proc Natl Acad Sci USA* **100**, 8589-94 (2003).
52. Silady, R.A. et al. The gravitropism defective 2 mutants of *Arabidopsis* are deficient in a protein implicated in endocytosis in *Caenorhabditis elegans*. *Plant Physiol* **136**, 3095-103 (2004).
53. Sedbrook, J.C. et al. ARG1 (altered response to gravity) encodes a DnaJ-like protein that potentially interacts with the cytoskeleton. *Proc Natl Acad Sci USA* **96**, 1140-5 (1999).
54. Harrison, B.R. & Masson, P.H. ARL2, ARG1 and PIN3 define a gravity signal transduction pathway in root statocytes. *Plant J* **53**, 380-92 (2008).
55. Morita, M.T. et al. A C2H2-type zinc finger protein, SGR5, is involved in early events of gravitropism in Arabidopsis inflorescence stems. *Plant J* **47**, 619-28 (2006).
56. Young, L.S. et al. Adenosine kinase modulates root gravitropism and cap morphogenesis in *Arabidopsis*. *Plant Physiol* **142**, 564-73 (2006).
57. Caspar, T. & Pickard, B.G. Gravitropism in a starchless mutant of *Arabidopsis*: Implications for the starch-statolith theory of gravity sensing. *Planta* **177**, 185-97 (1989).
58. Withers, J.C. et al. Gravity persistent signal 1 (GPS1) reveals novel cytochrome P450s involved in gravitropism. Am J Bot 100, 183-93 (2013).59.
59. Bennett, M.J. et al. *Arabidopsis* AUX1 gene: a permease-like regulator of root gravitropism. *Science* **273**, 948-50 (1996).
60. Swarup, R. et al. Structure-function analysis of the presumptive *Arabidopsis* auxin permease AUX1. *Plant Cell* **16**, 3069-83 (2004).
61. Friml, J. et al. Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*. *Nature* **415**, 806-9 (2002).
62. Noh, B. et al. Enhanced gravi- and phototropism in plant mdr mutants mislocalizing the auxin efflux protein PIN1. *Nature* **423**, 999-1002 (2003).