Dissecting the genetic variation and relationship of four botanical peanut varieties using whole chloroplast genome sequencing (#23303)

First submission

Editor guidance

Please submit by **2 Mar 2018** for the benefit of the authors (and your $200 publishing discount).

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Raw data check
Review the raw data. Download from the materials page.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the materials page.

6 Figure file(s)
3 Table file(s)
4 Other file(s)

Custom checks

DNA data checks

- Have you checked the authors data deposition statement?
- Can you access the deposited data?
- Has the data been deposited correctly?
- Is the deposition information noted in the manuscript?

Structure and Criteria

Structure your review

The review form is divided into 5 sections.
Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- ✓ Clear, unambiguous, professional English language used throughout.
- ✓ Intro & background to show context. Literature well referenced & relevant.
- ✓ Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- ✓ Figures are relevant, high quality, well labelled & described.
- ✓ Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- ✓ Original primary research within Scope of the journal.
- ✓ Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- ✓ Rigorous investigation performed to a high technical & ethical standard.
- ✓ Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- ℹ Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- ✓ Data is robust, statistically sound, & controlled.
- ✓ Conclusions are well stated, linked to original research question & limited to supporting results.
- ✓ Speculation is welcome, but should be identified as such.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Dissecting the genetic variation and relationship of four botanical peanut varieties using whole chloroplast genome sequencing

Juan Wang [1] , **Chunjuan Li** [1] , **Caixia Yan** [1] , **Xiaobo Zhao** [1] , **Shihua Shan** Corresp. [1]

[1] Shandong Peanut Research Institute, Qingdao, Shandong, China

Corresponding Author: Shihua Shan
Email address: shansh1971@163.com

Background: Since chloroplast is maternal transmission and non-recombination, the sequences have been used broadly in the taxonomic classification and phylogeny reconstruction. *Arachis hypogaea* L. is worldwide significant oilseed and economic crop. The complete chloroplast (cp) nucleotide sequences of four representative botanical varieties were obtained by next-generation sequencing (NGS).

Methods: To reveal their genome structures and phylogenetic relationship, the entire sequencing reads of var. *hypogaea* (AHP), var. *hirsuta* (AHL), var. *fastigiata* (AHD) and var. *vulgaris* (AHZ) were separately assembled and annotated. According to the alignment sequences, the genome-wide genetic variations (SNPs and InDels) were developed.

Results: The complete length of cp genome for AHP, AHL, AHD and AHZ was 156,354bp, 156,878bp, 156,718bp and 156,399bp, respectively. Comparative genome sequences analysis of the four types indicated that gene content, gene order and GC content were quite similar to each other, and a total of 97.8% SNPs and 88.5% InDels harbored in the non-coding regions. The phylogenetic relationships among the four botanical varieties suggested that AHL constituted a basal branch of the peanut group, which coincided with the previous records. Meanwhile, a higher variable region (*trnI*-GAU intron) was detected which is suitable for evolutionary studies at the intraspecific level.

Discussion: The four cp genome resources will provided valuable genetic message for accurately distinguishing cultivars and constructing the genetic relationship.

Dissecting the genetic variation and relationship of four botanical peanut varieties using whole chloroplast genome sequencing

Juan Wang[1*], Chunjuan Li[1*], Caixia Yan[1], Xiaobo Zhao[1], Shihua Shan[1#]

[1]Laboratory of Genetics and Breeding, Shandong Peanut Research Institute, Qingdao 266100, Shandong Province, China

[*]Juan Wang and Chunjuan Li have contributed equally to this work.

[#]corresponding author:

Shihua Shan

E-mail: shansh1971@163.com

Abstract

Background: Since chloroplast is maternal transmission and non-recombination, the sequences have been used broadly in the taxonomic classification and phylogeny reconstruction. *Arachis hypogaea* L. is worldwide significant oilseed and economic crop. The complete chloroplast (cp) nucleotide sequences of four representative botanical varieties were obtained by next-generation sequencing (NGS).

Methods: To reveal their genome structures and phylogenetic relationship, the entire sequencing reads of var. *hypogaea* (AHP), var. *hirsuta* (AHL), var. *fastigiata* (AHD) and var. *vulgaris* (AHZ) were separately assembled and annotated. According to the alignment sequences, the genome-wide genetic variations (SNPs and InDels) were developed.

Results: The complete length of cp genome for AHP, AHL, AHD and AHZ was 156,354bp, 156,878bp, 156,718bp and 156,399bp, respectively. Comparative genome sequences analysis of the four types indicated that gene content, gene order and GC content were quite similar to each other, and a total of 97.8% SNPs and 88.5% InDels harbored in the non-coding regions. The phylogenetic relationships among the four botanical varieties suggested that AHL constituted a basal branch of the peanut group, which coincided with the previous records. Meanwhile, a higher variable region (*trnI*-GAU intron) was detected which is suitable for evolutionary studies at the intraspecific level.

Discussion: The four cp genome resources will provided valuable genetic message for accurately distinguishing cultivars and constructing the genetic relationship.

Subjects Genomics, Plant Science

Keywords Peanut cultivars, Chloroplast genomes, Genetic variation, Genetic relationship

Short title: cp genome analysis of peanut cultivars

Introduction

Cultivated peanut (*Arachis hypogaea* L.) is an AABB-type (2n=4x=40) polyploidy species originated from South America after the relatively complicated evolutionary progress involving natural and artificial selection (Bertioli et al. 2016). Peanut, one of the essential oilseed crops, is mainly planted in China, India, American and Argentina (Hammons 1994; Grabiele et al. 2012). By the morphological observations, a large number of landraces were classified into four botanical varieties: variety (var.) *hypogaea* and var. *hirsute* belong to subspecies (ssp.) *hypogaea*, and var. *fastigiata* and var. *vulgaris* to ssp. *fastigiata* (Gibbons et al. 1972). Then, Krapovickas and Vanni (1960; 2010) added another two region-specific botanical varieties into ssp. *fastigiata* (var. *aequatoriana* and var. *peruviana*). The phenotypic characteristics of the cultivars are usually influenced by external factors. Thus, the phylogenetic relationship of these cultivars revealed by the molecular markers is more reliable than the traditional empirical method (Gepts 1993; He & Prakash 2001).

Compared with the nuclear sequence, the chloroplast (cp) sequence has its advantages including non-recombination, haploid and maternal inheritance (Birky 2001). The cpDNA has been often used for identifying species and dissecting phylogenetic relationships (Zhao et al. 2015; Jansen et al. 2007). For example, Grabiele et al. (2012) investigated the polymorphisms of two cultivated peanut subspecies using non-coding cpDNA regions (*trnTR-trnS* and *trnT-trnY*) and a non-transcribed spacer of the nuclear 5S rDNA markers. Although the result strongly indicated that the six botanical varieties had a single genetic origin, the phylogenetic relationship between these varieties was not illustrated because of limited sequence information.

With the development of sequencing technologies, the cp genomes of *Nicotiana tabacum* and *Marchantia polymorpha* were first reported (Ohyama et al. 1986; Shinozaki et al. 1986). Over the last few years, the cost-efficient genome data output largely benefit from the rapid progress of next generation sequencing (NGS). Prabhudas et al. (2016) reported the first cp genome sequences of *A. hypogaea*. The general features of *A. hypogaea* cp genome and genome structure dynamics have been well-described, which provided an ideal reference genome. The cp genomes are powerful for the accurately scanning DNA polymorphisms and effective in providing valuable inter-specific information for the reconstruction of phylogeny (Jansen et al. 2007; Parks et al. 2009; Moore et al. 2010). For example, Yin et al. (2017) developed seven species cp genomic resources of *Arachis* and provided the best resolution in molecular phylogeny. Besides, the cp genomes were also helpful for dynamic structure study at the subspecies level. For instance, Zhao et al. (2015) reported four Chinese *Panax ginseng* strains and found the identical cp genomes. Meanwhile, the minor allele sites indicated the cp genome was undergoing dynamic change to fit different environments.

As an important economic crop, A. *hypogaea* has been planted in China for more than 500 years, where has become the largest producer in the world (Yu 2008). These four botanical varieties, var. *hypogaea* (AHP), var. *hirsuta* (AHL), var. *fastigiata* (AHD) and var. *vulgaris* (AHZ) were already widely distributed in China. Given the genome data were insufficient for detection the variation (SNPs and InDels) and genetic relationship between the peanut cultivars, we developed four cp genome complete nucleotide sequences using high-throughput sequencing method in this study. Then we investigated the genetic relationships based on four peanut cultivars and other published genomes. Our results will supply more molecular resources for further variety identification and phylogenetic resolutions.

Materials & Methods

DNA extraction and sequencing

Four botanical varieties (*A. hypogaea* var. *hypogaea*, *hirsuta*, *fastigiata* and *vulgaris*) were collected from Shandong Peanut Research Institute, Qingdao, China. The seedlings were grown using hydroponic methods. Fresh leaves (> 5g) collected from the 3~4 weeks plant were used to isolate chloroplast DNA using Plant Chloroplast DNAOUT Kit (Bjbalb, China). The library with an average length of 350bp was constructed using NexteraXT DNA Library Preparation Kit (Illumina, China). The library quality was testified by GeneRead DNA QuantiMIZE Assay Kit (QIAGEN, Germany). Sequencing was performed on Hiseq Xten platform. The average length of the generated reads was 150 bp (Illumina, China).

Data assembly and annotation

The quality of the raw paired-end reads was assessed by FastQC v0.11.3 (Andrews 2014). All raw HiSeq data of four varieties was filtered based on the following rules: 1) adapter trimming; 2) reads quality control with <5% unidentified nucleotides and > 50% bases quality value >20. This work was accomplished using Cutadapt v1.7.1 (Martin 2011). Then, the high-quality data were used to *de novo* assembly (http://soap.genomics.org.cn; Luo et al. 2012). The assembled data were arranged according to the complete cp genome of *A. hypogaea* L. Co7 variety using Mauve v2.3.1 tool (Darling et al. 2010; Prabhudas et al. 2016). The cp genes were annotated by DOGMA tool with default parameters (Wyman et al. 2004). Genome pictures were drawn with OGDraw v1.2 (Lohse et al. 2007).

Variation detection and phylogenetic analysis

Multiple alignments were generated using VISTA and Mauve algorithm software v2.3.1 (Frazer et al. 2004; Darling et al. 2010) and checked manually. All alignments and related information were visualized using the VISTA viewer (Mayor et al. 2000). For retrieving InDels (insertions /deletions), the multiple alignment file was input MOSAIK (Lee et al. 2014; http://gkno.me/pipelines.html#mosaik). SSRs were separated from all filtered InDels. The phylogeny was constructed based on the whole genome sequences comprising IR (A/B) and (L/S)SC regions of peanut cultivars and other relative species. The close relative species of Fabaceae with high similarities (E value $<10^{-6}$) were regarded as outgroups. The phylogenetic tree was constructed by minimum evolution (ME) algorithm in MEGA v6 with default parameters (Tamura et al. 2011).

Results

Genome assembly and validation

High-throughput sequencing based on the Illumina Hiseq Xten system generated raw data (> 1G sequencing data per sample). After cleaning and trimming, 22,511,400 (AHZ) to 62,087,400 (AHL) paired-end reads were mapped separately to the reference cp genome reaching 143× to 396× coverage. After *de novo* and reference-guided assembly with minor modifications, we obtained four complete cp genome sequences (Figure 1; Table 1).

According to the assembled cp genome sequences, the .sqn files were separately generated using sequin software (https://www.ncbi.nlm.nih.gov/projects/Sequin/), submitted to NCBI Genbank and acquired the accession numbers: MG814006 (AHD); MG814007 (AHL); MG814008 (AHP); MG814009 (AHZ).

Size and gene content of the peanut genome

Among these four cp genomes, sequence length ranged from 156,354 bp to 156,878 bp. The size varied from 85,900 bp (AHL) to 86,196 bp (AHD) in the LSC region, from 18,796 bp (AHP, AHL and AHZ) to 18,874 bp (AHD) in SSC region and from 25,806 bp (AHP) to 26,091 bp (AHL) in IR (A/B) region (Table 1). A total of 110 unique genes harbored in cp genome in which containing four ribosomal RNA (rRNA) genes, 76 protein-coding genes and 30 transfer RNA (tRNA) genes (Table 2). Among these genes, 16 genes (Six of the protein-coding genes, six of the tRNA genes and four of the rRNA genes) were completely repeated in the IR(A/B), giving a total of 126 genes. The genome contained 55.66% coding regions and 44.34% noncoding regions, including both intergenic spacers and introns. The overall GC content of the cp sequence was 36.3~36.4% and the GC content for LSC, SSC, and IR(A/B) was 33.8%, 30.2~30.3%, and 42.8~42.9% respectively (Figure 2; Table 2).

DNA Flexibility

The flexible value of peanut cp genome was ranged from 9.87 to 12.21 (Figure 2). The higher flexible regions (top 5%) with maximum value of 12.21 were detected, including *psbK*-*accD* intergenic spacer (56131-57150), *trnL*-UAA*trnT*-UGU intron (14201-15280) and *ndhL* (120641-

121680). These regions were the start sites of RNA polymerase combination or transcription in favor of protein complex recognition. Meanwhile, the lower flexible regions (top 5%) with minimum value of 9.85 comprised two 23s ribosomal RNA blocks (108681-109690; 134081-135080), perhaps because of the requirement for base pairing in the secondary structures of the products.

Genome variations

The multiple alignments of peanut cp genome sequences were performed. All regions of the four peanut cultivars presented no differences in the junction positions (Figure 3). VISTA-based identity plots illustrated the hotspot regions of genetic variation between cp genomes (Figure 4). A total of 46 SNPs were found within the quadripartite region. As expected, non-coding regions harbored the higher variation than coding regions, and the higher substitutions were located in the *trnI*-GAU intron (25 SNPs) and *ycf3-psaA* spacer (8 SNPs) regions.

The total number of 26 InDels was detected: 13 in spacers, 9 in introns of genes and 4 in genes with 15 in LSC region, 2 in SSC region, and 9 in IRA /IRB regions (Supplementary Figure S1). Large InDels (>50 bp) were found in the *psbK-trnQ* intergenic spacer, *trnL* intron (IR), *ycf1* among the four botanical varieties. Among them, we identified 6 SSR regions with >7 repeat nucleotides with sequence identify >90%: 4 A stretches and 1 T stretches ranging from 7 bp to 16 bp, and 1 with dinucleotide repeat motifs of CTAG. No C or G stretches were identified.

Phylogenetic analysis

According to the similarity result, *Robinia pseudoacacia*, *Ceratonia silique*, *Leucaena trichandra* and *Senna tora* of Fabaceae were used as outgroups. Due to the low genetic diversity, whole genome sequences were used to construct the phylogenetic tree based on ME algorithms. The result showed the six genome sequences of peanut cultivars were clustered into a monophyletic branch. AHL constituted a basal clade compared with other peanut cultivars (Figure 5). AHZ was close to AHL; then were the *A. hypogaea* KX257487 and KJ468094. AHD and AHP were clustered together. Meanwhile, other species were grouped into the other group. The high support values (> 99%) were shown above nodes.

Discussion

The chloroplast (cp) is a cyclic organelle in plant cytoplasm originated from cyanobacteria. The chloroplast was in charge of photosynthesis and carbon fixation (Alberts et al. 2002). The chloroplast usually lack recombination and was maternally inherited, which makes it an important reference for understanding the phylogenetic and taxon distinguishing. Here, we compared the whole cp genome sequences for AHP, AHL, AHZ and AHD based on NGS method and revealed the divergence of the entire cp genome. All four complete peanut cp genomes displayed the classic quadripartite structure. There were no obvious genomic rearrangements and gene inversion. Comparative genomic sequences indicated that gene content and gene order of these four types were well-conserved as expected.

Non-synonymous variations

The highest variation number (25 of 46 SNPs) was identified in *trnI*-GAU intron region, which could provide fruitful information for the variety identification, and can be used to generate useful DNA barcode for *Arachis*. Most substitutions and InDels were synonymous. Only one substitution in *psaA* gene was involved in nonsynonymous mutation. The *psaA* gene is a fundamental protein-coding gene of photosystem I. The hydrophobic amino acid Tyr of *psaA* gene in AHD, AHZ and AHP was replaced by a hydrophilic amino acid Asn in AHL, which indicated that AHL may develop a modified photosystem I to adjust their ability to adapt to the changing photosynthetic environment during the domestication process (Wu et al. 2017). Besides, three InDels in IRA *ycf1* and IRA /IRB *ycf2* regions had resulted in protein functional change. Specifically, the 63 bp-insertion at the end of *ycf1* gene led to a longer amino acid sequence in AHD, while a 18 bp-deletions was found in the middle of IRA /IRB *ycf2* gene in AHP. The *ycf1* gene has recently been re-recognized as a crucial protein component of the cp translocon located at the inner envelope membrane (Kikuchi et al. 2013). The 63 bp-tail in AHD may acquire additional function for cp translocon. The *ycf2* gene is the largest plastid gene in plants. Huang et al. (2010) showed that the *ycf2* gene alone could provide a consistent and well-supported phylogenetic relationship instead of the most gene combinations. While in peanut, the genome-wide variations were easier to distinguish the botanical varieties.

The earliest domesticated cultivar

Six available genome sequences of peanut cultivars and the additional genome resources of Fabaceae were employed in the study of phylogenetic relationships. The varieties belonging to a ssp. *hypogaea* or *fastigiata* were mixed together. It is possible that these four varieties were closely related by the maternal transmission. Combined the nuclear sequence information, they may lead to a better disclosing of the entire phylogenetic process. However, the phylogenetic performed successfully addressed the following evolutionary issue. Our results suggested that AHL constituted a basal branch compared with other cultivars and had a close phylogenetic relationship to other species of Fabaceae, which were in good accordance with previous reports. AHL is the most similar to wild species morphologically (Krapovickas et al. 1960). More importantly, AHL is regarded as the earliest peanut cultivar that was domesticated in the South American based on the historical record. And then, AHL was introduced in China where is now considered as a secondary differentiation center (Krapovickas et al. 1960; Duan et al. 1995). Thus, AHL was considered as an ancient botanical variety, which was supported by our molecular evidence.

Conclusion

We reported four complete cp genomes of peanut cultivars using Illumina sequencing methods. The gene contents and gene orders of cp genomes were showed highly conserved. We investigated the genetic variations (SNPs and InDels) of the four complete peanut cp genomes. The non-coding regions, *trnI*-GAU intron region was considered as rapidly-evolving regions that could be the potential molecular maker for the phylogenetic study. Moreover, our results raise more evidence to support the hypothesis that AHL is the ancient variety of the peanut cultivars. This study was a better attempt to unseal high-supported phylogenetic relationship of cultivated peanut.

Acknowledgements

The authors would like to acknowledge Dr. Dachuan Shi from Qingdao Academy of Agricultural Science for the excellent advice on this paper.

Funding

This work was supported by the Natural Science Foundation of Shandong Province (grant number ZR2017BC082); the Specific Funds of the Central Guidance for Local Science and Technology; the International Science & Technology Cooperation Program of China (grant number 2015DFA31190).

References

Alberts B, Johnson A, Lewis J, et al. 2002. *Molecular Biology of the Cell*. 4th edition. Chapter 14: Chloroplasts and Photosynthesis. New York: Garland Science.

Andrews S. 2014. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EK, Liu X, Gao D, Clevenger J, Dash S, Ren L, Moretzsohn MC, Shirasawa K, Huang W, Vidigal B, Abernathy B, Chu Y, Niederhuth CE, Umale P, Araujo AC, Kozik A, Kim KD, Burow MD, Varshney RK, Wang X, Zhang X, Barkley N, Guimaraes PM, Isobe S, Guo B, Liao B, Stalker HT, Schmitz RJ, Scheffler BE, Leal-Bertioli SC, Xun X, Jackson SA, Michelmore R, and Ozias-Akins P. 2016. The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nat Genet* 48:438-446.

10.1038/ng.3517

Birky CW. 2001. The Inheritance of Genes in Mitochondria and Chloroplasts: Laws, Mechanisms, and Models. *Annual Review of Genetics* 35:125-148.

Darling AE, Mau B, and Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147. 10.1371/journal.pone.0011147

Duan NX, Jiang HF, Liao BS, Zhou R. 1995. var. *hirsuta* in China: the origin and spread (in Chinese). Chinese Journal of Oil Crop Sciences 17 (2): 68-71.

Frazer KA, Pachter L, Poliakov A, Rubin EM, and Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32:W273-279. 10.1093/nar/gkh458

Gepts P. 1993. The Use of Molecular and Biochemical Markers in Crop Evolution Studies. *Evolutionary Biology-new York* 27:51-94.

Gibbons RW, Bunting AH, and Smartt J. 1972. The classification of varieties of groundnut (Arachis hypogaea L.). *Euphytica* 21:78-85.

Grabiele M, Chalup L, Robledo G, and Seijo G. 2012. Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. *Plant Systematics and Evolution* 298:1151-1165.

Hammous RO. 1994. The origin and history of the groundnut. In: The groundnut crop: a scientific basis for improvement. Chapman and Hall, New York. DOI: 10.1007/978-94-011-0733-4_2

He G, and Prakash C. 2001. Evaluation of genetic relationships among botanical varieties of cultivated peanut (Arachis hypogaea L.) using AFLP markers. *Genetic Resources and Crop Evolution* 48:347-352. 10.1023/a:1012019600318

Huang J, Sun G, and Zhang D. 2010. Molecular evolution and phylogeny of the angiosperm ycf2 gene. *Journal of Systematics and Evolution* 48:240-248.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee SB, Peery R, McNeal JR, Kuehl JV, and Boore JL. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci U S A* 104:19369-19374. 10.1073/pnas.0709121104

Kikuchi S, Bedard J, Hirano M, Hirabayashi Y, Oishi M, Imai M, Takase M, Ide T, and Nakai M. 2013. Uncovering the Protein Translocon at the Chloroplast Inner Envelope Membrane. *Science* 339:571-574.

Krapovickas et al. 1960. revista de investigaciones agricolas 14(2): 197-228.

Krapovickas A and Gregory WC. 1994. Taxonomia del énero *Arachis* (Leguminosae). *Bonplandia* 8: 1–186. DOI: 10.2307/41941177.

Krapovickas A, and Gregory WC. 2010. TAXONOMY OF THE GENUS ARACHIS (LEGUMINOSAE) Taxonomy of the genus Arachis (Leguminosae). *Bonplandia*.

Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, and Marth GT. 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One* 9:e90581. 10.1371/journal.pone.0090581

Lohse M, Drechsel O, and Bock R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet* 52:267-274. 10.1007/s00294-007-0161-y

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, and Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18. 10.1186/2047-217X-1-18

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* 17. 10.14806/ej.17.1.200

pp. 10-12

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, and Dubchak I. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16:1046-1047.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, and Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci U S A* 107:4623-4628. 10.1073/pnas.0907801107

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, and Chang Z. 1986. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature* 322:572-574.

Parks M, Cronn R, and Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* 7:84-84.

Prabhudas SK, Prayaga S, Madasamy P, and Natarajan P. 2016. Shallow Whole Genome Sequencing for the Assembly of Complete Chloroplast Genome Sequence of Arachis hypogaea L. *Front Plant Sci* 7:1106. 10.3389/fpls.2016.01106

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayshida N, Matsubayasha T, Zaita N, Chunwongse J, Obokata J, and Yamaguchishinozaki K. 1986. The complete nucleotide sequence of the tobacco chloroplast genome. *Plant Molecular Biology Reporter* 4:111-148.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731-2739. 10.1093/molbev/msr121

Wyman SK, Jansen RK, and Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252-3255. 10.1093/bioinformatics/bth352

Wu XP, Sen L, Chen N, Zhang X, et al． 2017. Study on the molecular evolution of the *psaA* gene from ferns． Plant Science Journal 35(2): 177-185. 10. 11913 /PSJ. 2095-0837.

Yin D, Wang Y, Zhang X, Ma X, He X, and Zhang J. 2017. Development of chloroplast genome resources for peanut (Arachis hypogaea L.) and other species of Arachis. *Sci Rep* 7:11649. 10.1038/s41598-017-12026-x

Yu SL. 2008. Peanut varieties and their genealogy in China. Shanghai Science and Technology

Press.

Zhao Y, Yin J, Guo H, Zhang Y, Xiao W, Sun CM, Wu J, Qu X, Yu J, and Wang X. 2015. The complete chloroplast genome provides insight into the evolution and polymorphism of Panax ginseng. *Frontiers in Plant Science* 5:696-696.

Figure 1

Gene map of the *A. hypogaea* chloroplast genomes

Genes shown outside the outer circle are transcribed clockwise and those inside are transcribed counterclockwise. Genes belonging to different functional groups are color-coded. Dashed area in the inner circle indicates the GC content of the chloroplast genome.

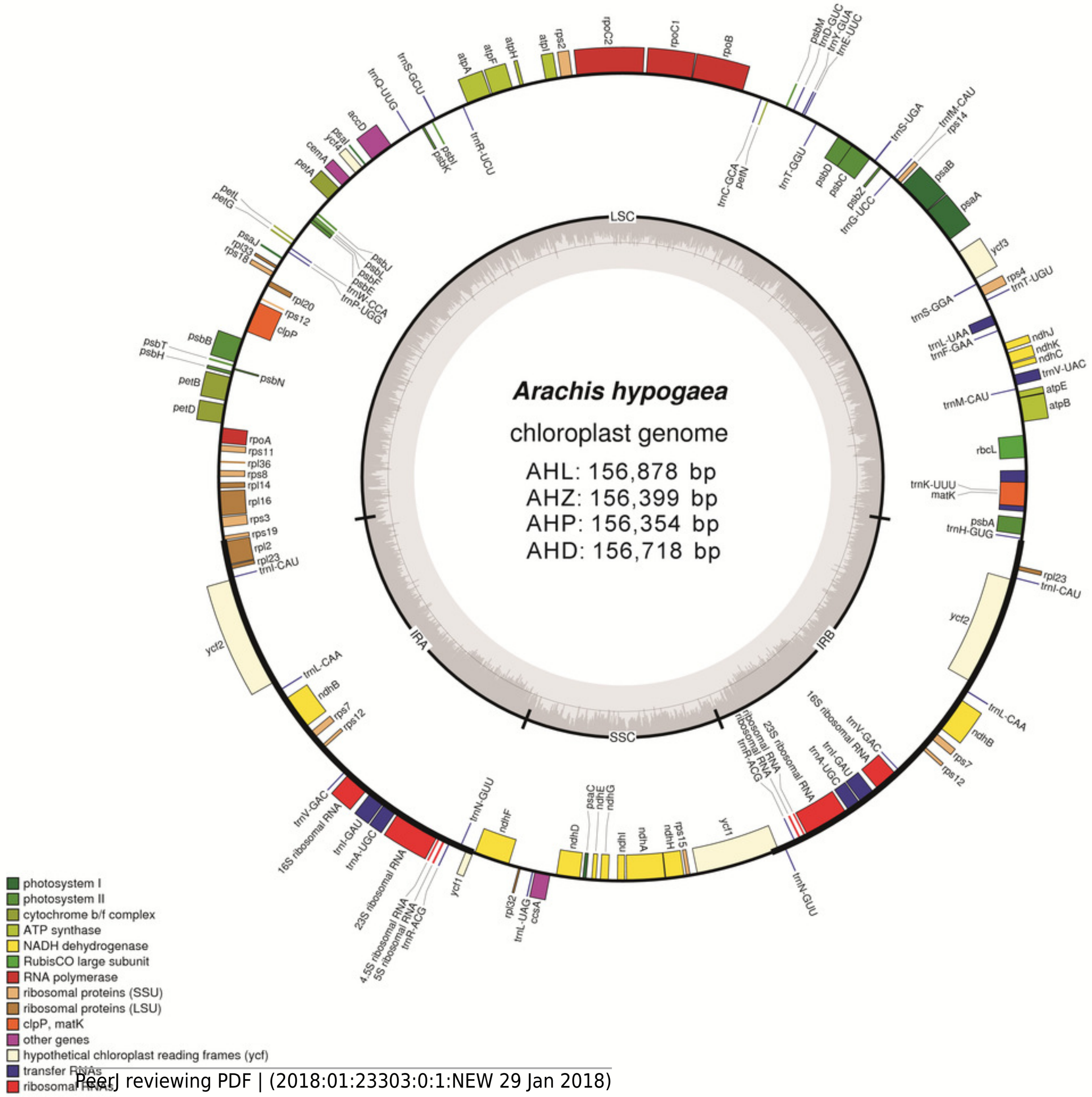

Figure 2

DNA helix flexibility ranged from 9.87 to 12.21 were shown in the upper graph

Plot of G+C and A+T composition along cp genome using 10 sliding windows (1000 bp per window). Accordingly, the genome composition is being shown in the bottom graph. Four corresponded regions were plotted by non-gray and gray highlighting.

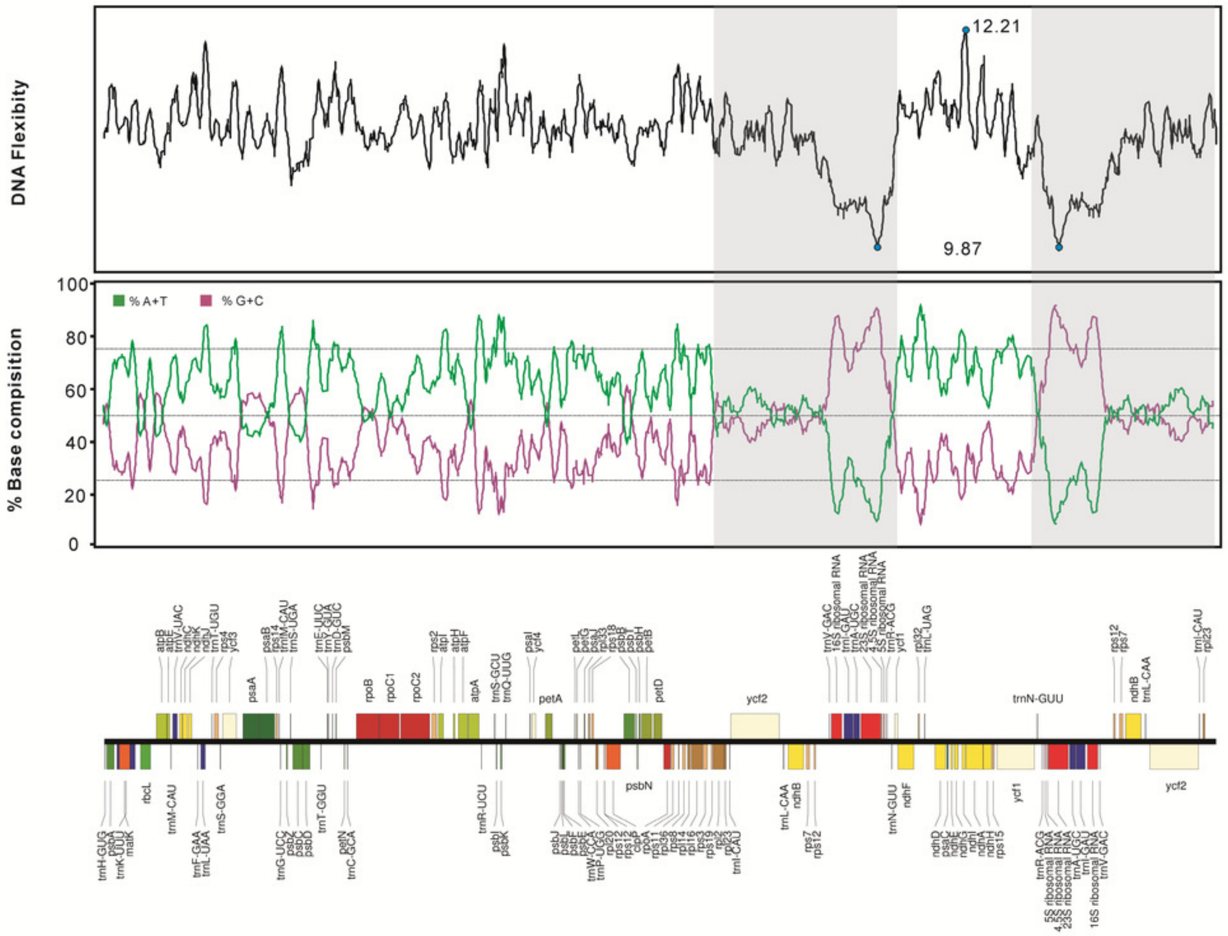

Figure 3

The comparison of the LSC, IR and SSC border regions among the four peanut chloroplast genomes

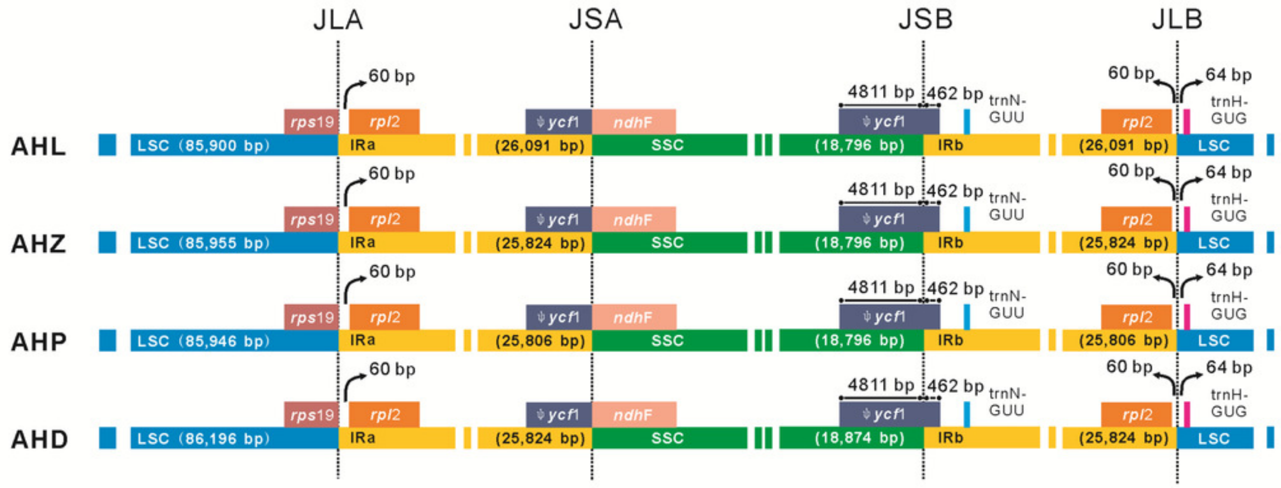

Figure 4

Visualization of alignment of the peanut chloroplast genome sequences

Genome regions are color-coded as protein coding, rRNA coding, tRNA coding or conserved noncoding sequences (CNS). The x-axis represents the coordinate in the chloroplast genome. Annotated genes are displayed along the top.The sequences similarity of the aligned regions is shown as horizontal bars indicating the average percent identity between 50% and 100%.

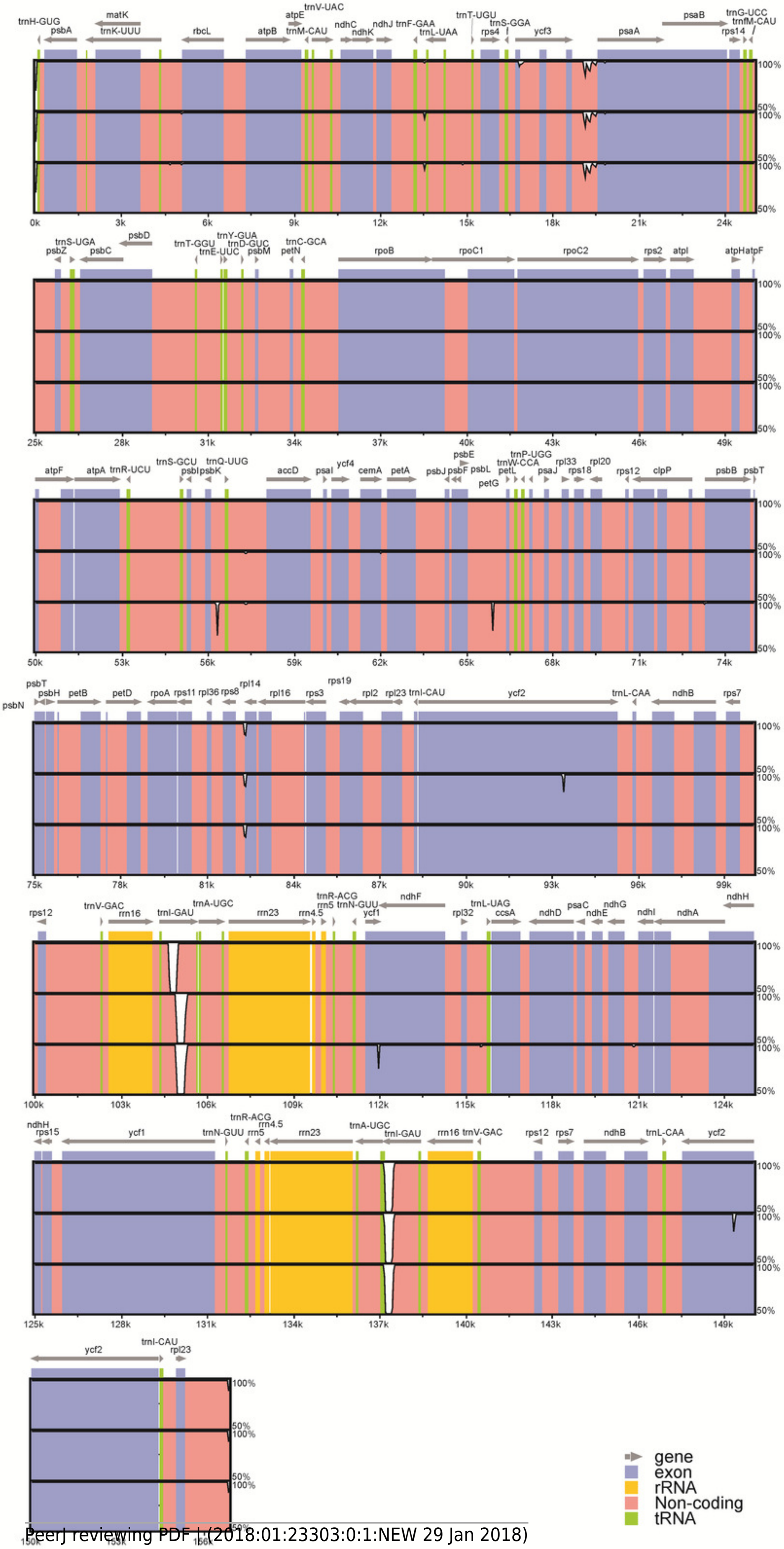
trnH-GUG
psbA
matK
trnK-UUU
rbcL
atpB
atpE
trnV-UAC
trnM-CAU
ndhC
ndhK
ndhJ
trnF-GAA
trnL-UAA
trnT-UGU
rps4
trnS-GGA
ycf3
psaA
psaB
trnG-UCC
trnfM-CAU
rps14
0k
3k
6k
9k
12k
15k
18k
21k
24k
100%
50%
trnS-UGA
psbZ
psbC
psbD
trnT-GGU
trnE-UUC
trnY-GUA
trnD-GUC
psbM
petN
trnC-GCA
rpoB
rpoC1
rpoC2
rps2
atpI
atpH
atpF
25k
28k
31k
34k
37k
40k
43k
46k
49k
atpF
atpA
trnR-UCU
trnS-GCU
psbI
psbK
trnQ-UUG
accD
psaI
ycf4
cemA
petA
psbJ
psbE
psbF
psbL
petG
petL
trnW-CCA
trnP-UGG
psaJ
rpl33
rps18
rpl20
rps12
clpP
psbB
psbT
50k
53k
56k
59k
62k
65k
68k
71k
74k
psbT
psbH
psbN
petB
petD
rpoA
rps11
rpl36
rps8
rpl14
rpl16
rps3
rps19
rpl2
rpl23
trnI-CAU
ycf2
trnL-CAA
ndhB
rps7
75k
78k
81k
84k
87k
90k
93k
96k
99k
rps12
trnV-GAC
rrn16
trnI-GAU
trnA-UGC
rrn23
rrn4.5
trnR-ACG
rrn5
trnN-GUU
ycf1
ndhF
rpl32
trnL-UAG
ccsA
ndhD
psaC
ndhE
ndhG
ndhI
ndhA
ndhH
100k
103k
106k
109k
112k
115k
118k
121k
124k
ndhH
rps15
ycf1
trnN-GUU
trnR-ACG
rrn4.5
rrn5
rrn23
trnA-UGC
trnI-GAU
rrn16
trnV-GAC
rps12
rps7
ndhB
trnL-CAA
ycf2
125k
128k
131k
134k
137k
140k
143k
146k
149k
ycf2
trnI-CAU
rpl23
150k
153k
156k
gene
exon
rRNA
Non-coding
tRNA

Figure 5

The evolutionary relationship among four cultivated peanuts and the related species of Fabaceae constructed by NJ analyses

Numbers above node are bootstrap support values.

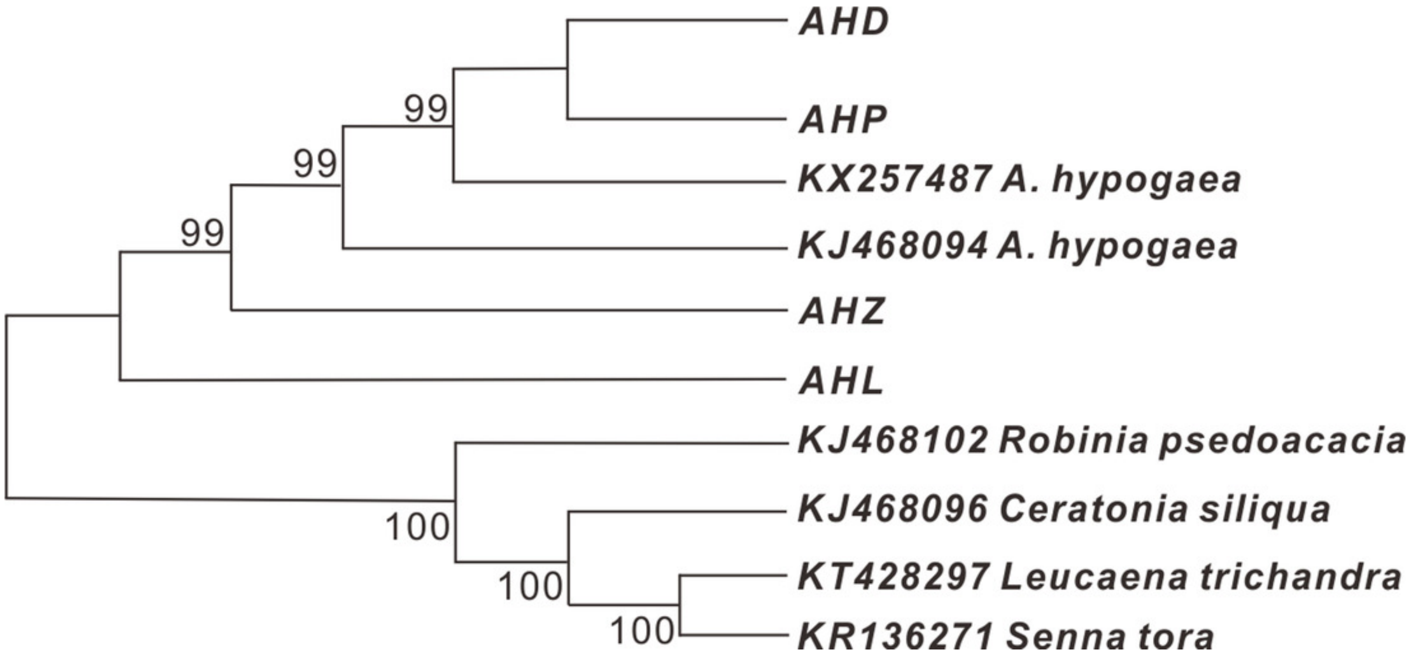

Table 1(on next page)

Genes identified in the chloroplast genome of peanut

Intron-containing genes are marked by asterisks (*).

Table 1 Genes identified in the chloroplast genome of peanut. Intron-containing genes are marked by asterisks (*).

Category for genes	**Group of genes**	**Name of genes**
Self-replication	tRNA genes	*rrn5, rrn4.5, rrn16, rrn23*
	rRNA genes	**trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, *trnG-UCC, trnH-GUG, trnI-CAU, *trnI-GAU,*trnK-UUU, trnL-CAA, *trnL-UAA, trnL-UAG, trnfM-CAU,trnM-CAU, trnN-GUU, trnP-UGG, trnQ-UUG,trnR-ACG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU,trnT-UGU, trnV-GAC, *trnV-UAC, trnW-CCA, trnY-GUA*
	small subunit of ribosome	*rps2, rps3, rps4, rps7, rps8, rps11, *rps12, rps14,rps15, *rps16, rps18, rps19*
	large subunit of ribosome	*rpl2, rpl14, *rpl16, rpl20, rpl22, rpl23, rpl32, rpl33,rpl36*
	DNA dependent RNA polymerase	*rpoA, rpoB, *rpoC1, rpoC2*
Genes for photosynthesis	Subunits of NADH-dehydrogenase	**ndhA, *ndhB, ndhC, ndhD, ndhE, ndhF,ndhG, ndhH, ndhI, ndhJ, ndhK*
	Subunits of photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbN, psbT, psbZ*
	Subunits of cytochrome b/f complex	*petA, *petB, *petD, petG, petL, petN,*
	Subunits of ATP synthase	*atpA, atpB, atpE, *atpF, atpH, atpI*
	Large subunit of rubisco	*rbcL*
Other genes	Maturase	*matK*
	Protease	**clpP*
	Envelope membrane protein	*cemA*
	Subunit of Acetyl-CoA-carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
Genes of unknown function	Open Reading Frames (ORF, ycf)	*ycf1, ycf2,*ycf3,ycf4*

Table 2(on next page)

Details of the complete chloroplast genomes of four peanut botanical varieties

Table 2 Details of the complete chloroplast genomes of four peanut botanical varieties.

	AHL	**AHZ**	**AHP**	**AHD**
Matched reads (bp)	62,087,400	22,511,400	61,928,100	34,570,200
Genome size (bp)	156,878	156,399	156,354	156,718
Mean coverage(×)	395.77	143.94	396.08	220.59
LSC length (bp)	85,900	85,955	85,946	86,196
SSC length (bp)	18,796	18,796	18,796	18,874
IR length (bp)	26,091	25,824	25,806	25,824
LSC GC content (%)	33.8	33.8	33.8	33.8
SSC GC content (%)	42.9	42.9	42.9	42.9
IR GC content (%)	30.3	30.3	30.3	30.2
GC content (%)	36.4	36.4	36.4	36.3
Total	110	110	110	110
Protein coding genes	76	76	76	76
rRNA	4	4	4	4
tRNA	30	30	30	30