~~C~~A comparative analysis of the complete chloroplast genome sequences of four peanut botanical ~~types~~varieties

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, China

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

[#]corresponding author:
Shihua Shan

E-mail: shansh1971@163.com

Abstract

Background: *Arachis hypogaea* L. is an economically important ~~–~~oilseed crop worldwide ~~and there are~~ comprising six botanical varieties ~~within the species. There are still limited chloroplast (cp) genomic resources available for this species.~~ ~~Here~~In this work, we ~~investigated~~ characterized the chloroplast (cp) genome sequences of the four widely distributed peanut varieties. ~~The complete chloroplast (cp) genome sequences of four representative botanical types were obtained by next-generation sequencing (NGS); these sequences will provide useful genomic resources for further phylogeny reconstruction.~~

Methods: The cp genome data of four widely distributed botanical varieties (var. *hypogaea*, var. *hirsuta*, var. *fastigiata* and var. *vulgaris*) were obtained by next-generation sequencing (NGS). These high throughput sequencing ~~data~~ reads ~~of the four studied *A. hypogeae* types (var. *hypogaea*, *hirsuta*, *fastigiata* and *vulgari*)~~were then assembled, annotated and comparatively analyzed.

Results: The total cp genome lengths of ~~–~~the studied *A. hypogaea* varieties were~~was~~ 156,354 bp (~~for~~ var. *hypogaea*), 156,878 bp (~~for~~ var. *hirsuta*), 156,718 bp (~~for~~ var. *fastigiata*~~-~~) and 156,399 bp (~~for~~ var. *vulgaris*)~~, respectively~~. Comparative ~~cp genome sequence~~ analysis of theses cp genome sequences~~of these four types~~ revealed that their gene content, gene order and GC content were ~~quite similar to each other~~highly conserved, with only a total of 46 SNPs and 26 InDels identified ~~among them~~. Most of the variation is restricted to non-coding sequences, especially, the *trnI*-GAU intron region~~. In addition, a~~ was detected to be highly ~~variable~~ variable~~region (*trnI* GAU intron) was also detected which~~ and will be useful for future evolutionary studies.~~–~~

Discussion: The~~se~~ four cp genome sequences acquired here will provide valuable genetic resources for distinguishing ~~the four studied~~ *A. hypogaea* botanical varieties~~types~~ and determining their evolutionary relationship.

Short title:~~–~~

Peanut cp genome~~s~~ analysis

Introduction

Cultivated peanut (*Arachis hypogaea* L.) is one of the most important oilseed crops that is mainly planted in China, India, the United States of America and Argentina (Hammons 1994; Grabiele et al. 2012; Bertioli et al. 2016). Based on morphological (Gibbons et al. 1972, Krapovickas and Vanni, 1960; 2010) and molecular (Gepts 1993; He & Prakash 2001) evidences, six botanical varieties of *A. hypogaea* have been identified: var. *hypogaea*, var. *hirsuta*, var. *fastigiata*, var. *vulgaris* (Gibbons et al. 1972), as well as var. *aequatoriana* and var. *peruviana*, with the last two being region specific.

In land plants, the cp genome is circular and has a large single copy (LSC) region and a small single copy (SSC) region that are separated by a pair of inverted repeats (IRs) regions (Raubeson and Jansen 2005). The major role of the chloroplast is to conduct photosynthesis; additionally, it is also involved in the biosynthesis of fatty acids, vitamins, pigments and amino acids (Prabhudas et al. 2016). Different from nuclear sequence, the cp DNA has several advantages, including low-recombination, haploid ploidy, and maternal inheritance, making it an ideal tool for evolutionary studies (Birky 2001; Wu and Ge 2012). For example, with the help of genetic markers that includes non-coding cpDNA regions (*trnTR-trnS* and *trnT-trnY*), Grabiele et al. (2012) found that the six peanut botanical varieties were very likely to have a single genetic origin, however, the fine evolutionary relationship between these varieties remains to be resolved.

The rapid progress of high-throughput sequencing technology development has greatly facilitated the acquisition of cp genome data, which are not only powerful for reconstructing interspecific

phylogeny (Jansen et al. 2007; Parks et al. 2009; Moore et al. 2010), but are also helpful for investigating genome dynamic ~~structure study~~ at the subspecies level. For instance, Zhao et al. (2015) ~~reported~~ compared the cp genomes of four Chinese *Panax ginseng* strains and ~~found the minor allele sites, which indicated the cp genome was undergoing dynamic change to fit different environments.~~suggested that their genome dynamic was under selective pressure.

~~China has become the largest producer of cultivated peanuts in the world (Yu 2008), and the four *A. hypogaea* botanical types in this study, has been planted there for more than 500 years.~~ Although there are six botanical varieties within *A. hypogaea* that differ at both the morphological and molecular levels (Ferguson et al. 2004), only very limited *A. hypogaea* cp genome data are currently available (Prabhudas et al. 2016; Choi and Choi 2017). Here, we acquired and ~~closely~~ examined the ~~four~~ complete cp genome ~~complete~~ nucleotide sequences of the four main peanut botanical varieties ~~the availability of more peanut cp genomes will supply more molecular~~, providing valuable genetic resources for further evolutionary studies~~resolutions~~.

Materials & Methods

DNA extraction and sequencing

Four representative *A. hypogaea* ~~botanical types~~varieties (var. *hypogaea*, var. *hirsuta*~~,~~ var. ~~var. ff~~*astigiata* and var. *vulgaris*) were collected from Shandong Peanut Research Institute, Qingdao, China. China has become the largest producer of cultivated peanut in the world (Yu 2008), and these four main botanical varieties have been cultivated in China for more than 500 years. The seedlings were grown using hydroponic methods. The cp DNA was isolated from fresh leaves (> 5 g) ~~collected from~~of 3~~~~ to 4- week-old seedlings using the Plant Chloroplast DNAOUT Kit (Bjbalb, China). The ~~Q~~quality of cp DNA samples ~~were~~ was ~~detected~~ checked by agarose gel electrophoresis with Super GelRed (US Everbright Inc, Suzhou, China)~~.~~. ~~The l~~Libraries with an average length of 350 bp ~~was~~ were constructed using ~~a~~ the NexteraXT DNA Library Preparation Kit (Illumina, China). The quality of the libraries ~~were tested~~was checked ~~—~~by

Formatted: Font: Italic

Formatted: Font: Italic

GeneRead DNA QuantiMIZE Assay Kit (QIAGEN, Germany). Sequencing was performed on the Illumina HiSeq Xten platform (Illumina, China), and 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~~ for four ~~–~~ *A. hypogaea* varieties were filtered based on the following rules: 1) adapter trimming; 2) ~~reads~~ quality control ~~with~~; each read has < 5% unidentified nucleotides and > 50% of its bases with a quality value of > 20. This filtration was ~~accomplished~~ carried out using Cutadapt v1.7.1 (Martin 2011). The high-quality data were then assembled into contigs using the *de novo* assembler SPAdes v3.9.0 (Nurk et al. 2013), and these contigs were further assembled into complete cp genome~~s by further connection~~ using NOVOPlasty (Dierckxsens et al. 2016)~~.~~. The assembled data were checked against ~~the~~ the published complete cp genome of *A. hypogaea* ~~var. Co7~~ (GenBank accession no. KX257487, Prabhudas et al. 2016). The cp genes were annotated ~~by~~ using the DOGMA tool with ~~the~~ default parameters (Wyman et al. 2004). The cp genome images were drawn with OGDraw v1.2 (Lohse et al. 2007).~~–~~

Variation detection ~~of variation~~ and ~~analysis of~~ evolutionary relationship analysis~~–~~

Multiple sequence alignment ~~were~~ was generated using VISTA and Mauve v2.3.1 software (Frazer et al. 2004; Darling et al. 2010) and ~~were~~ was checked manually when necessary. All alignments ~~and related information~~ were visualized using the VISTA viewer program (Mayor et al. 2000). Single nucleotide polymorphisms (SNPs) were ~~detected~~ identified by Mauve ~~software~~ v2.3.1. ~~Insertions~~The insertions/deletions (InDels) were retrieved from the sequence alignment ~~sequence~~s using the mVISTA package. An InDels image including 10 bp up- and downstream was then generated. Simple sequence repeats (SSRs) were isolated from all filtered InDels. Repeat sequences with repeating units of 2-6 bp ~~and~~ that repeated no fewer than three times were considered as SSRs. ~~DNA flexibility was used to target the regions of high DNA~~

~~helix flexibility within la DNA sequence calculated by the Graphs package of Unipro UGENE (Okonechnikov et al. 2012).~~

~~Genetic~~ The genetic relationship of the four peanut cp genomes together with two available peanut cp genome sequences (GenBank accession no. KX257487 and KJ468094; Prabhudas et al. 2016; Choi and Choi 2017) were examined by constructing a minimum evolutionary (ME) tree using MEGA v6 with default parameters (Tamura et al. 2011). ~~revealed based on the whole cp genome sequences of the peanut botanical types and other related species~~ The cp genome sequences from four other related species (*Robinia pseudoacacia*, *Ceratonia siliqua*, *Leucaena trichandra* and *Senna tora*) of Fabaceae were used as outgroups (CSI-BLAST E-value < 10^{-6}). ~~The closely related species of the Fabaceae with a high similarity (E value <10^{-6}) were considered outgroups. A minimumevolutionary tree was constructed via the minimum evolution (ME) algorithm in MEGA v6 with the default parameters (Tamura et al. 2011).~~

Formatted: Superscript

Results

~~cp genome assembly~~Assembly and validation of cp genomes

~~High throughput sequencing based on the Illumina HiSeq Xten system generated raw data (> 1Gb sequencing data per sample)~~More than 1 Gb raw sequencing data per sample was generated from high-throughput sequencing. After cleaning and trimming, 22,511,400 (var. *vulgaris*~~.~~) to 62,087,400 (var. *hirsuta*) paired-end reads were acquired, which were then mapped separately to the reference cp genome, attaining coverage amounts of 143× to 396×. After performing ~~the~~ *de novo* and reference-guided assembly with minor modifications, we acquired four complete cp genome sequences ~~that belong to, respectively~~ for~~,~~ *A. hypogaea* ~~varieties~~ var. *hypogaea*~~,~~ var. *hirsuta*, var. *fastigiata*~~,~~ ~~var. *fastigiata*~~ and var. *vulgaris* (Figure 1; Table 1).

For each of the assembled cp genome sequences, a .sqn file that was generated ~~using~~ by the ~~sequin~~ Sequin software ~~—~~(https://www.ncbi.nlm.nih.gov/projects/Sequin/), submitted to ~~NCBI Genbank~~ GenBank and acquired the following accession numbers: MG814006 ~~(~~for var. *fastigiata*~~)~~, MG814007 ~~(~~for var. *hirsuta*~~)~~, MG814008 ~~(~~for var. *hypogaea*~~),~~ and MG814009 for ~~(~~var. *vulgaris*~~)~~. Users can download the data for research

purposes only ~~with quoting the~~when referencing this paper.

Genetic structure of the peanut cp genome

These four acquired peanut cp genomes were found to have the classical quadripartite structure of land plant chloroplast genomes that comprises a LSC, a SSC and two IR (A/B) regions. The sequence lengths among the four cp genomes ranged from 156,354 bp to 156,878 bp. The size varied from 85,900 bp (var. *hirsuta*) to 86,196 bp (var. *fastigiata*) in the LSC region, from 18,796 bp (var. *hypogaea*, var. *hirsuta* and var. *vulgaris*) to 18,874 bp (var. *fastigiata*) in the SSC region and from 25,806 bp (var. *hypogaea*) to 26,091 bp (var. *hirsuta*) in the IR (A/B) region (Table 1). A total of 110 ~~unique~~ genes ~~in the cp genome contained~~ were identified from the cp genome: four ribosomal RNA (rRNA) genes, 76 protein-coding genes and 30 transfer RNA (tRNA) genes (Table 2). Among the 110 identified genes, ~~Six~~ six protein-coding genes, six tRNA genes and four rRNA genes were distributed in the IR (A/B) regions. ~~–~~The cp genome consisted of 55.66% ~~of~~ coding regions and 44.34% ~~of~~ non-coding regions~~,~~ including both intergenic spacers and introns. The overall GC content of the cp genomic sequences was 36.3~36.4%, and the GC contents ~~in~~ of the LSC, SSC, and IR (A/B) regions ~~was~~ were 33.8%, 30.2~30.3%, and 42.8~42.9%, ~~–~~respectively (~~Figure 2;~~ Table 2).

~~DNA flexibility~~

~~The flexibility value of the peanut cp genome ranged from 9.87 to 12.21 (Figure 2). Regions of higher flexibility (top 5%) with a maximum value of 12.21 were detected, including the *psbK-accD* intergenic spacer region (56,131-57,150 bp), *trnL*-UAA*trnT*-UGU intron (14,201-15,280 bp) and *ndhL* (120,641-121,680 bp). These regions were the start sites of the RNA polymerase complex or the transcription site – for protein complex recognition. Moreover, the regions of less flexibility (top 5%) with a minimum value of 9.85 comprised two 23S rRNA blocks (108,681-109,690 bp; 134,081-135,080 bp), perhaps because of the requirement for base pairing in the secondary structures of the products.~~

Variation among the cp genomes ~~variations~~

~~Multiple alignments of the peanut cp genome sequences were performed. N~~Among the four acquired peanut cp genome sequences, there was~~e regions in the four peanut botanical types presented~~ no ~~differed~~ difference at the junction positions (Figure ~~3~~2). ~~VISTA-based identity plots illustrated the hotspot regions of genetic variation among the cp genomes (Figure 4).~~ A total of 46 SNPs were found within the quadripartite structural region. VISTA-based identity plots illustrated the hotspot regions of genetic variation among the cp genomes (Figure 43). As expected, non-coding ~~regions~~ sequences exhibited more variation than did the coding sequences~~regions~~, and ~~the~~ greater amounts of substitutions were found in the *trnI*-GAU intron (25 SNPs) and the *ycf3-psaA* spacer (8 SNPs) regions. The only identified non-synonymous was located within the *psa*A gene. The hydrophobic amino acid Tyr in var. *hypogaea*, var. *fastigiata* and var. *yulgaris* was replaced by the hydrophilic amino acid Asn in var. *hirsuta*.

A total of 26 InDels were ~~detected~~identified: ~~13~~ thirteen were located in spacers, ~~9~~ nine were in introns, and ~~4~~ four were in genes; 15 were in the LSC region, ~~2~~ two were in the SSC region, and ~~9~~ nine were in IR (A /–B) regions (Supplementary Figure S1). Among these InDels~~four botanical varieties~~, large InDels (> 50 bp) were found in the *psbK-trnQ* intergenic spacer, the *trnL* intron ~~(an IR),~~ and *ycf1*. Among those InDels, we identified ~~6~~ six SSR regions ~~with >7 repeat nucleotides whose~~ (sequence identity ~~was~~ > 90%): ~~4~~ four A stretches and ~~1~~ one T stretches ranging from 7 bp to 16 bp, as well as ~~1~~ one with a CTAG ~~dinucleotide~~ repeat motif. No C or G stretches were identified. Moreover, InDels in the *ycf1* and the *ycf2* regions represent frameshift mutations: ~~Specifically,~~ the 63 bp-insertion at the end of the *ycf1* gene led to a longer amino acid sequence in var. *fastigiata*, while a 18 bp-deletion~~s~~ was found in the middle of IR (A /B) *ycf2* gene regions in var. *hypogaea*.

Genetic relationship analysis

~~The similarity results showed that *Robinia pseudoacacia*, *Ceratonia silique*, *Leucaena trichandra* and *Senna tora* of the Fabaceaeserved as outgroups.~~ Due to ~~the~~ low genetic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

diversity, the whole cp genome sequences were used to construct an evolutionary tree based on ME algorithms. The results showed that these peanut cp genomes ~~sequences of the six peanut varieties~~ clustered into a monophyletic branch, while the four outgroup species were clustered into another branch. Among the six analyzed peanut cp genomes, var. *hirsuta* is relatively different from the rest and constitute a basal clade (Figure 4). ~~Compared with other peanut types, var. *hirsuta* constituted a basal clade (Figure 5). Four peanut types were clustered together. Meanwhile, other species were grouped into the other group.~~ The high support values (> 99%) were shown above the nodes.

Discussion

The chloroplast (cp) is an important plant cell organelle ~~peculiar to plant cytoplasm originated from cyanobacteria~~ (Alberts et al. 2002). The ~~chloroplast~~ cp genome usually lacks recombination and are is maternally inherited~~, as such, they represent an important reference for understanding the phylogenetic relationships and~~and is therefore very useful for distinguishing taxa and inferring evolutionary relationships. Here, we have studied the cp genomes of cultivated peanut (*A. hypogaea*) that is an economically important oilseed crop worldwide. *A. hypogaea* comprised six varieties that differ at both the morphological and molecular levels (Ferguson et al. 2004). So far, only very limited *A. hypogaea* cp genome data are available (Prabhudas et al. 2016). In the present study, we acquired and closely examined the whole cp genome sequences of four main peanut varieties. We found that the overall cp genome structures of the four botanical varieties were the same and displayed the classical quadripartite structure of land plant cp genome (Raubeson and Jansen 2005).~~we compared the whole cp genome sequences for var. *hypogaea*, *hirsuta*, *fastigiata* and *vulgaris* based on NGS methods and revealed the variation within the entire cp genome.~~ No definitive genomic rearrangements or gene inversions were found among the four peanut cp genomes. The sequence variation among the four peanut cp genomes was also relatively limited, and most of them were restricted to the noncoding regions, especially the *trnI*-GAU intron exhibited an outstanding level of variation (25 out of the entire 46

Formatted: Font: Italic

identified SNPs), suggesting that the rapidly evolving nature of this intron. This *trnI*-GAU intron has therefore a great potential for developing molecular markers that could be used in future phylogenetic studies.

All four complete peanut cp genomes displayed the classic quadripartite structure. There were no definitive genomic rearrangements or gene inversions. A Comparison of the genomic sequences indicated that gene content and gene order among these four types were well conserved, as expected.

Formatted: Font: Font color: Red

Non-synonymous variations

The greatest frequency of variation (25 of 46 SNPs) was identified in the *trnI* GAU intron region, which could provide useful information for the variety identification, and can be used to generate useful DNA barcodes 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 the *psaA* gene in var. *hypogaea*, *fastigiata* and *vulgaris* was replaced by a hydrophilic amino acid Asn in var.*hirsuta*, which indicated that *hirsuta* may have evolved a modified photosystem I to adjust their ability to adapt to the changing photosynthesis environment during its domestication process (Wu et al. 2017). In addition, The *ycf1* gene product has recently been re-recognized as a crucial component of the cp translocon located at the inner envelope membrane (Kikuchi et al. 2013). The 63 bp tail in *fastigiata* may have acquired additional function for the cp translocon. The *ycf2* gene is the largest plastid gene in plants. Huang et al. (2010) reported that the *ycf2* gene alone could provide a consistent and well-supported phylogenetic relationship instead of most gene combinations. In peanut, the cp genome-wide variationscould easily distinguish the botanical varieties.

Genetic relationships among the botanical types

TIn addition, a minimum-evolution tree of the four acquired peanut cp genomes together with two earlier published peanut cp genomes has been constructed to speculate their evolutionary relationships. he available cp genome sequences of six

~~peanut varieties and the additional cp genome resources of the Fabaceae were used into study evolutionary relationships.~~ Our result showed that the six investigated peanut cp genomes form a monophyletic branch, and this agrees with earlier studies (Grabiele et al. 2012). In addition, our result also revealed that among the six studied peanut cp genomes, var. *hirsuta* was relatively more distantly related to the others and may constitute a basal branch, which was in line with the previous reports (Duan et al. 1995; Ferguson et al. 2004). Consistent with its suggested relationship between var. *hirsuta* and the other studied peanut varieties, var. *hirsuta* appeared to be the peanut variety found within the archeological remains along the Pacific coast of Perú (Bonavia) that may be the region of origin of cultivated peanut (Simpson et al 2001; Stalker et al. 2017).~~The varieties belonging to subspecies *hypogaea* or *fastigiata* were mixed together. These four varieties were possibly closely related by the maternal transmission. Combined with their nuclear sequence information, these types could lead to a better understanding of the entire evolutionary process. However, Our results suggested that, compared with other varieties, var. *hirsuta* constituted a basal branch, which was in good accordance with previous reports (Ferguson et al. 2004). var. *hirsuta* is the variety found in the most ancient archeological remains along the Pacific coast of Perú (Bonavia) (Simpson et al 2001; Stalker et al. 2017). var. *hirsuta* was considered as the ancient cultivated type in China, where has become a secondary center of diversity (Duan et al. 1995).).~~

Conclusion

~~Using Illumina sequencing methods~~With the help of high-throughput sequencing technology, we revealed the complete cp genomes of four main peanut botanical varieties~~types~~. The gene contents and gene orders of the cp genomes were highly conserved. ~~We investigated the genetic variations among the four complete peanut cp genomes.~~ The ~~noncoding regions and the~~ *trnI*-GAU intron region was considered to be rapid~~ly~~-evolving region~~s~~ that could ~~be~~ potentially serve as molecular markers ~~for~~ in phylogenetic studies. ~~Moreover, our results provide more evidence to support the hypothesis that var. *hirsuta* is the relatively ancient ancient botanical type.~~ This study

will provide ~~more~~ valuable cp ~~genome~~ genomic resources for ~~the phylogeny reconstruction of *A. hypogaea*~~future exploitation.

Acknowledgements

The authors would like to acknowledge Dr. Yuan Li from Lund University in Sweden and Dr. Dachuan Shi from ~~the~~ Qingdao Academy of Agricultural Science in China for their excellent advices on this paper.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, and Watson JD. 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. https://doi.org/citeulike-article-id:11583827

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. *Nature Genetics* 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.

Choi IS, and Choi BH. 2017. The distinct plastid genome structure of *Maackia fauriei* (Fabaceae: Papilionoideae) and its systematic implications for genistoids and tribe Sophoreae. *PLoS One* 12(4):e0173766. doi:10.1371/journal.pone.0173766

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

Dierckxsens N, Mardulyn P, and Smits G. 2017. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45: e18. 10.1093/nar/gkw955

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

Ferguson ME, Bramel PJ, and Chandra S. 2004. Gene Diversity among Botanical Varieties in Peanut (L.). *Crop Science* 44 (5): 1847-1854.

Frazer KA, Pachter L, Poliakov A, Rubin EM, and Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Research* 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(1):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. 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 C, 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. *Proceedings of the National Academy of*

Sciences of the United States of America 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 A, and Rigoni VA. 1960. La nomenclatura de las subespecies y variedades de *Arachis hypogaea* L. *Revista De Investigaciones Agricolas "Alberto Boerger"* 14: 198–228.

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

Krapovickas A, ~~,~~ and Gregory WC. 2007. Taxonomy of the genus *Arachis* (Leguminosae). *Bonplandia*. 16: 7-205.

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. *Current Genetics* 52:267-274. 10.1007/s00294-007-0161-y

Katoh K, and Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology & Evolution*. 30(4):772-780.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet Journal* 17(1). 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. *Proceedings of the National Academy of Sciences of the United States of America* 107:4623-4628. 10.1073/pnas.0907801107

Formatted: Portuguese (Portugal)

Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, McLean J, Lasken R, Clingenpeel SR, Woyke T, Tesler G, Alekseyev MA, and Pevzner PA. 2013. Assembling genomes and mini-metagenomes from highly chimeric reads. *International Conference on Research in Computational Molecular Biology* (Vol. 7821, pp.158-170). Springer, Berlin, Heidelberg.

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.

Okonechnikov K, Golosova O, and Fursov M, the UGENE team. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28: 1166-1167. 10.1093/bioinformatics/bts091

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. *Frontier in Plant Science* 7:1106. 10.3389/fpls.2016.01106

Raubeson LA, and Jansen RK. 2005. Chloroplast genomes of plants," in *Plant diversity and evolution: genotypic and phenotypic variation in higher plants*, ed Henry RJ. (Wallingford, CT: CABI), 45-68.

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.

Simpson CE. 2001. Use of wild *Arachis* species/ introgression of genes in to *A. hypogaea* L. *Peanut Science* 28(2): 114-116.

Stalker HT. 2017. Utilizing wild species for peanut improvement. *Crop Science* 57(3):

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

1102-1120. 10.2135/cropsci2016.09.0824

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. *Molecular Biology and Evolution* 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 ZQ, and Ge S. 2012. Phylogeny of the BEP clade in grasses revisited: evidence from whole genome sequences of chloroplast. *Molecular Phylogenetics and Evolution* 62: 578–578. doi: 10.1016/j.ympev.2011.10.019

Wu XP, Sen L, Chen N, Zhang X, Ma ZX, Zhang QY. 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*. *Scientific Reports* 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.

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: EndNote Bibliography, Indent: Left: 0 cm, Hanging: 1,27 cm