Insights into amphicarpy from the compact genome of the legume *Amphicarpaea edgeworthii*

Yiyang Liu1,†, Xuejie Zhang2,†, Kai Han3,†, Rongchong Li1, Guoxin Xu4, Yan Han2, Feng Cui3, Shoujin Fan2, Inge Seim5, Guangyi Fan3,6,7,8, Guowei Li1,* and Shubo Wan1,†

1Bio-technology Research Center, Shandong Provincial Key Laboratory of Crop Genetic Improvement, Ecology and Physiology, Shandong Academy of Agricultural Sciences, Ji’nan, China
2College of Life Sciences, Shandong Normal University, Ji’nan, China
3BGI-Qingdao, BGI-Shenzhen, Qingdao, China
4Shandong Rice Research Institute, Shandong Academy of Agricultural Sciences, Ji’nan, China
5Integrative Biology Laboratory, College of Life Sciences, Nanjing Normal University, Nanjing, China
6BGI-Shenzhen, Shenzhen, China
7State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, China

Received 2 October 2020; revised 13 November 2020; accepted 18 November 2020.
*Correspondence (Tel +86-531-66658127; fax:+86-531-66652303; email wannshubo2016@163.com (S.W.), Tel +86-531-66658902; fax:+86-531-66658156; email lgw_sas@163.com (G.L.), Tel +86-532-55711134; fax:+86-532-55711134; email fanguangyi@genomics.cn (G.F.)).
†These authors contribute equally to this work.

Keywords: Amphicarpaea edgeworthii, amphipcarpy, genome size, genome assembly, leguminous species.

Abstract

Amphicarp (seed heteromorphy) is a unique and fascinating reproductive strategy wherein a single plant produces both aerial and subterranean fruits. This strategy is believed to be an adaptation to life under stressful or uncertain environments. Here, we sequenced and de novo assembled a chromosome-level genome assembly of the legume *Amphicarpaea edgeworthii* Benth. The 299-Mb A. edgeworthii genome encodes 27,899 protein-coding genes and is the most compact sequenced legume genome reported until date. Its reduced genome size may be attributed to the reduced long-terminal repeat retrotransposon content, which stems from the unequal homologous recombination. Gene families related to immunity and stress resistance have been contracted in A. edgeworthii, which is consistent with the notion that the amphicarpic reproductive strategy may be a complementary mechanism for its weak environmental-adaptation ability. We demonstrated the ‘ABCE’ model for the differentiation of chasmogamous and cleistogamous flowers. In addition, the characteristics of aerial and subterranean seeds in hard-seededness were explored. Thus, we suggest that the A. edgeworthii genome, which is the first of an amphicarpic plant, offers significant insights into its unusual reproductive strategy that is a key resource towards comprehending the evolution of angiosperms.

Introduction

Angiosperms can be categorized into the following 4 types based on the spatial locations of their fruits: aerocarpic, basicarpic, geocarpic, and amphicarpic (van der Pijl, 1982). Geocarpic and amphicarpic are largely found in the terrestrial herbaceous plants inhabiting niche habitats that occupied are ephemeral or susceptible to stochastic disturbances (Ellner and Shmida, 1981). For the geocarpic species, all fruits develop underground, while the flowers are borne aboveground or underground (full geocarpic) (Barker, 2005). Amphicarpic literally means two types of seeds and it can be categorized into (i) amphicarpic sensu stricto that refers to plants producing aerial and subterranean flowers and aerocarpic and geocarpic fruits and (ii) amphicarpic sensu lato that refers to plants producing only aerial flowers, but both aerocarpic and geocarpic fruits (Zhang et al., 2015).

Amphicarp has been considered as an intermediate life-history strategy between aerocarp and geocarp (Sharma et al., 1981). Although geocarp and amphicarp are distributed across different taxa, there is considerable homoplasy in their morphological characteristics (Kaul et al., 2000). The fruiting modes are determined by the floral types, which are generally divided into chasmogamous (i.e., open; CH) and cleistogamous (i.e., closed; CL) flowers with distinct morphologies (Oakley et al., 2007). Amphicarpic is usually associated with cleistogamy, especially of the basal or underground flowers, and heterodiaspory (Barker, 2005). Based on the study of cleistogamous flower variations in *Viola* species in Hawaii (Ballard et al., 1998), it has been proposed that the ancestral *Viola* colonists may have been strictly CH and that dimorphic cleistogamy subsequently evolved, or the chasmogamy evolved from dimorphic cleistogamy as a result of the loss of CL flower (Culley and Klooster, 2007).

The occurrence is mostly frequent in the Fabaceae family with in 36 species with the amphicarpic fruits including *Trifolium polymorphum* and *Pisum fulvum* etc. (Zhang et al., 2020). *A. edgeworthii* also called as hogpeanut by native Americans, belonging to *Fabaceae Amphicarpaea*, is widely distributed across China and is a typical amphicarpic legume bearing both geocarpic and aerocarpic fruits in an individual plant (Figure 1a) (Zhang et al., 2005). It bears both geocarpic fruits with mostly 1-seeded pods and aerocarpic fruits with 2-4 seeded pods. *A. edgeworthii* can produce 3 types of flowers: aerial CH flowers (1 or 2
subtended by a bract); aerial CL flowers (generally solitary in leaf axils) and subterranean CL flowers (solitary without subtending leaves). The aerial CH flowers of *A. edgeworthii* differ significantly from the 2 types of CL flowers in the petal form, fertile stamen number, anther and carpel form, and ovule number, while the marked difference between 2 CL flower types is detected only in the ovule number (Zhang et al., 2015). CL flowers and amphicarpy have been proposed as a hierarchy of investments that is employed by small or stressed plants to allocate most of their energy to produce CL flowers and fruits or a 'pessimistic strategy' for adapting to uncertain environmental conditions (Schnie and Waller, 1986; Waller, 1980).

Until date, no genome of *Amphicarpaeae* availability has been a limiting factor for further studies on the evolution of amphicarpy. In the present work, we performed de novo assembly a chromosome-level genome of *A. edgeworthii* by combining long-read nanopore sequencing with the Hi-C scaffolding. The reference genome lays a foundation for genetics and evolution studies of *Amphicarpaeae*, which gives us an impetus to explore the differentiation mechanisms between flowers of CL and Ch for promotion of studies on the mechanisms of amphicarpy formation and evolution.

## Results

### Genome assembly and annotation

We generated 77.12-Gb clean data using the DNBSEQ sequencing platform and estimated the genome size of *A. edgeworthii* to be 319.65 Mb by the k-mer analysis (Figure S1). We identified approximately 52.28 Gb (174.27×) high-quality long-read nanopore data with an average read length of 25,051 bp (Figure S2; Table S1) and used them to assemble a genome of total size 299.04 Mb with a contig N50 length of 9.70 Mb (Table 1; Table S2). Using 208.27 million Hi-C reads (Table S3), 99.7% of the whole genome sequences (298.1 Mb) were successfully clustered, ordered, and oriented into 11 chromosomes (Figure S3 and Table S4); this outcome was consistent with the chromosome number identified by karyotype analyses (2n = 22) (Liang et al., 2009) (Figure 1b). The final assembly comprised of 24 scaffolds with a scaffold N50 length of 27.22 Mb, and the longest scaffold length was 30.61 Mb (Figure S4; Table S5). Based on this high-quality genome, we identified that transposable elements (TEs) occupied 40.18% (120.14 Mb) including 14.97% of long terminal repeats (LTRs) retrotransposon with a remarkable accumulation of Ty3-gypsy (11.78%) and Ty1-copia (6.86%) (Table 1). We also predicted 27,899 protein-coding genes in the *A. edgeworthii* genome by combining the approaches of *de novo*, homology searching, and transcriptome sequencing-based prediction methods (Table 1; Figure S5; Table S6). Approximately 98.5% (27,489) of protein-coding genes were functionally annotated (Table S7), and 13,201 miRNAs, 43,913 rRNAs, 53,920 tRNAs, and 43,546 snRNAs were also identified (Table 1; Table S8). The completeness was 98.0% and 92.2% when the BUSCO was run on the complete genome or gene sets, respectively (Tables S9 and S10).

### Phylogenetics and whole-genome duplications of *A. edgeworthii* genome

To compare the genome evolution in Leguminosae, we constructed a high-confidence phylogenetic tree and estimated the divergence times of 10 plant species using genes extracted from 1034 single-copy families. The phylogenetic tree revealed that *A. edgeworthii* belongs to the Millettioideae subfamily and that it is closely related to the soybean clade (*Glycine*), with a divergence time of 12.3 (6.4–17.9) million years ago (Mya) after the divergence of the common bean lineage 23 Mya (Figure 1b). We next detected the whole genome duplication (WGD) events for *A. edgeworthii* and other 5 Millettia species. The nucleotide diversity for the fourfold degenerate third-codon transversion site (4DTv) showed a shared peak (4DTv ~ 0.7) in the *A. edgeworthii* genome and five other Millettioideae species, which reflects that the whole-genome triplication event is common to the core eudicots and that another clear peak (4DTv ~ 0.30) was consistent with an ancestral Papilionoideae WGD event, which was estimated to have occurred approximately 55 Mya (Figure 1c). However, the *A. edgeworthii* genome did not experience any recent WGD event in *Glycine* (4DTv ~ 0.1), indicating that Glycine-specific duplication occurred after the divergence of *Glycine* and *A. edgeworthii* approximately 12.3 Mya (Figure 1c). The results from the distribution of synonymous substitutions per synonymous site (Ks) for *A. edgeworthii* paralog pairs were consistent with the results inferred from the 4DTv (Figure 1d).

To further investigate the genome evolution, we examined the collinear relationships between *A. edgeworthii* and common bean, adzuki bean, and soybean. One-to-one syntenic block pairwise relationship was observed between *A. edgeworthii* and common bean chromosomes; however, some chromosomes of *A. edgeworthii* matched to more than one chromosome of *P. vulgaris* (Figure 1e; Figure S6). For example, 2 segments in chromosome 1 (chr1-dup) and chromosome 5 (chr5-dup) matched with 2 syntenic blocks of C05 and C11 of common bean. Interestingly, some common chromosome fragments of *A. edgeworthii* and adzuki bean matched to more than one chromosome of common bean (Figure S7). These results suggest the occurrence of chromosome rearrangements, resulting in the deletion of paralogous genes in these genomes after shared WGD events.

### The evolution of the compact genome of *A. edgeworthii*

Genome size varies significantly among the embryophyta (Leitch et al., 2005). Interestingly, we noticed that the genome of *A. edgeworthii* is the smallest among the sequenced leguminous species. According to the angiosperm DNA C-values database, even among 920 Leguminosae species, only 5 have a smaller genome than that of *A. edgeworthii* (Bennett and Leitch, 2012). These species belong to the genera *Lotus* (3), *Leucaena* (1), and *Trifolium* (1) (Figure S8). On the other hand, *A. edgeworthii*, together with its close relatives in the Millettieae (such as common bean and adzuki bean), has experienced 2 common WGD events, which has resulted in genome expansion, but *A. edgeworthii* has the most compact genome. We next determined the cause of its small genome size. For this purpose, we compared protein-coding genes (non-redundant) in Leguminosae and found that *A. edgeworthii* (27,899 genes) has a similar gene repertoire to that of common bean and chickpea (Figure 2a; Table S11). Moreover, the average length of mRNA in *A. edgeworthii* is actually longer than that of common bean and others (Table S6). However, the genome size of common bean and chickpea is almost twice that of *A. edgeworthii*. It is well-documented that, in addition to WGD, TEs had obvious and significant impacts on the genome size during the evolution. Amplification, deletion, and rearrangements of repeated DNA sequences may account for intraspecific variations in genome size (Bennetzen et al., 2005). A comparison among 9 other

© 2020 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 19, 952–965
Figure 1 Morphology and evolution of A. edgeworthii genome. (A) Morphology of Amphicarpaea edgeworthii Benth. (a), aerial stem; (b), root; (c), subterranean flower; (d), subterranean fruit; (e), subterranean seed; (f), aerial CL flowers; (g), aerial CH flowers and (h), aerial pod and seeds; (b) The phylogeny, divergence time and orthogroup expansions / contractions for nine leguminous species and V. vinifera. The tree was constructed by maximum likelihood method using 1034 single copy orthogroups. Divergence time was estimated on a basis of seven calibration points (grey circles). The numbers in green and red indicate the numbers of orthogroups that have expanded and contracted beneath the species names, respectively; (c) Density distribution of 4DTv for paralogous genes; (d) Dot plot of chromosomal synteny (≥30 genes in each block) between A. edgeworthii and common bean chromosomes. The chromosome orientations are clockwise.
Table 1 Summary of the Amphicarpaea edgeworthii genome assembly and annotation

| Categories          | Type                  | Length (bp) | Number | Percentage (%) |
|---------------------|-----------------------|-------------|--------|----------------|
| Assembly            | Contigs               | 299,040,813 | 61     | –              |
|                     | Contig N50            | 9,697,753   | 12     | –              |
|                     | Contig N90            | 3,068,788   | 31     | –              |
|                     | Max length of contigs | 16,344,735  | –      | –              |
|                     | Scaffolds             | 299,059,313 | 24     | –              |
|                     | Scaffold N50          | 27,224,559  | 6      | –              |
|                     | Scaffold N90          | 25,196,224  | 10     | –              |
|                     | Max length of Scaffolds| 30,614,130   | –      | –              |
| Protein-coding genes| Total gene number     | –           | 27,899 | 100.0          |
|                     | Annotated gene        | –           | 27,489 | 98.5           |
|                     | Average length of mRNA| 3,777      | –      | –              |
|                     | Average length of CDS | 1,177      | –      | –              |
| Non-coding RNAs     | miRNA                 | 112         | 13,201 | 0.004          |
|                     | tRNA                  | 585         | 43,913 | 0.014          |
|                     | rRNA                  | 227         | 53,920 | 0.018          |
|                     | snRNA                 | 388         | 43,546 | 0.014          |
| Transposable elements| Class I: Retroelements| 87,996,628 | –      | 29.43          |
|                     | LTR/Copia             | 35,222,755  | –      | 11.78          |
|                     | LTR/Gypsy             | 20,505,147  | –      | 6.86           |
|                     | LINEs                 | 13,414,089  | –      | 4.49           |
|                     | SINEs                 | 111,672     | –      | 0.04           |
|                     | Class II: DNA Transponsors| 16,498,974 | –      | 5.52           |
|                     | Satellites            | 661,876     | –      | 0.22           |
|                     | Simple repeat         | 1,456,728   | –      | 0.49           |
|                     | Unknown               | 20,669,629  | –      | 6.91           |
|                     | Total TEs             | 120,141,897 | –      | 40.18          |

Insights into amphicarpy from the genome of *A. edgeworthii*

**Table 1** Summary of the Amphicarpaea edgeworthii genome assembly and annotation

Leguminous species revealed that the proportion of TEs varies greatly from 33.96% in *L. japonicus* and 40.45% in *A. edgeworthii* to 65.55% in soybean and 72.09% in common bean. Moreover, the total length of TEs of *A. edgeworthii* genome is shorter than that of the other species investigated, except for *L. japonicus*. We thus surveyed the correlation between genome size and the size of different TE classes (Table S12). The total TE length showed a strong correlation with genome size (*r*^2^ = 0.93, Figure 2b). Among the TE super families, Ty3-gypsy and Ty1-copia showed positive correlation with the genome size (*r*^2^ = 0.91 and *r*^2^ = 0.65, respectively; Figure 3b). These results indicate that the loss or reduced long-terminal repeat (LTR) retrotransposons contribute to the genome size of *A. edgeworthii*. We identified 978 Ty3-gypsy and 2444 Ty1-copia elements from full-length LTR retrotransposons (LTR-RTs) and accordingly constructed phylogenetic trees with soybean, common bean, and adzuki bean (Table S13). The members of all retrotransposon clades in soybean, common bean, and adzuki bean were also identified in *A. edgeworthii*. However, the copy number of both Ty3-gypsy and Ty1-copia in each clade in *A. edgeworthii* was consistently less than that in others (red lines; Figure 2c,d). To determine the historical dynamics of the LTR-RTs in the *A. edgeworthii* genome, we surveyed the insertion times of intact LTR elements. The majority of intact LTRs of Ty3-gypsy and Ty1-copia were relatively young, and most of the intact LTR elements were inserted within 1.0 Mya in soybean, common bean, and adzuki bean, suggesting that most elements became active only recently. However, both the copy number and the proportion of young LTR elements were lower in *A. edgeworthii* than that in the other genomes. Especially, the proportion of Ty3-gypsy had decreased dramatically most recently (approximately 0.1 Mya), which suggests that the transposition activity of *A. edgeworthii* was inhibited and the LTR elements were fragmented or deleted (Figure 2e). Rapid changes in the genome size were driven by bursts of LTR retrotransposon proliferation, followed by aggressive deletion through unequal homologous recombination and double-strand break repair (Bennetzen and Wang, 2014). Therefore, solo LTRs are a common outcome of misalignment and unequal crossing over between 2 LTRs on separate chromosomes or in a single LTR (Devos et al., 2002). We hence suspected that intensive and rapid removal of LTR-RTs was one of the causes of the apparent decrease in *A. edgeworthii* transposon activity (Bennetzen and Wang, 2014). In agreement, we found numerous solo LTRs in the *A. edgeworthii* genome. Importantly, *A. edgeworthii* has a much higher ratio (5.38) of solo/intact LTR elements as compared to soybean and adzuki bean (4.56 and 3.80, respectively; Table S14), indicating that large-scale recombination deletions have occurred in *A. edgeworthii* during evolution. Interestingly, functional annotation of species-specific gene families revealed that *A. edgeworthii* had gained families associated with homologous recombination and nucleotide excision repair (Table S15).

**Mass loss of stress-related genes**

The expansion and contraction of gene families in plants play major roles in plant phenotypic diversification in response to the selection pressure during evolution (Demuth and Hahn, 2009). To gain an insight into specific differences in the genic repertoire of *A. edgeworthii* and their potential biological significance, we compared the gene families in the *A. edgeworthii* and other 9
species. As a result, 30 expanded (P < 0.05) and 215 contracted (P < 0.05) gene families were identified in *A. edgeworthii*. The KEGG and Gene Ontology (GO) enrichment analyses indicated that the expanded gene families were mostly enriched in the pathways of cutin, suberin and wax biosynthesis, fatty acid biosynthesis, and oxidative phosphorylation, and in the GO terms of NADH dehydrogenase activity, carboxypeptidase activity, and transition metal ion binding (Tables S16 and S17). For the contracted gene families, the top 3 significantly enriched families included plant-pathogen interaction, linoleic acid metabolism, and glycosaminoglycan degradation in the KEGG pathways, and protein kinase activity, ATP binding, and catalytic activity in the GO terms, respectively (Tables S18 and S19).

The formation of CL flowers and amphicarpy may adapt to adverse environments (Cheplick, 1987; Culley and Klooster, 2007). Notably, among the contracted gene families, we identified that the gene number of well-studied resistance gene families related to immunity and stresses were decreased in *A. edgeworthii* (Figure 3a), which included LRR-RLK, NBS-LRR, serine/threonine (S/T) protein kinase, H⁺-transporting ATPases,
heat-shock protein network genes, and multicopper oxidases. Leucine-rich repeat receptor-like protein kinases (LRR-RLKs) are the largest group of receptor-like kinases in plants that play crucial roles in plant development and stress responses (Liu et al., 2017). A. edgeworthii has 174 LRR-RLK genes, while 194-475 RLKs are encoded in the other 8 legume species. Similarly, A. edgeworthii harbours only 62 S/T protein kinases, which are required for adaptation to biotic and abiotic stresses, while the other legume species possesses 103 to 255 genes. The heat-shock protein network was involved in conferring multiple stress resistance, especially in high-temperature stress and pathogen infection (Jacob et al., 2017). We detected only a few heat-shock proteins network-related isoforms which included 4 HSP70, 3 HSP90, and 2 HSFs in A. edgeworthii. The gene losses in the HSPs/HSF families may be one of the reasons why A. edgeworthii lives under the humid and shady conditions. The multicopper oxidases are copper-containing oxidoreductase enzymes that demonstrate exceptional performance as catalysts for the oxygen reduction reaction (McCaig et al., 2005). There are only 4 multicotpper oxidase genes in A. edgeworthii genome in comparison to 14 to 27 genes in the other 8 legume species. In plants, the nucleotide-binding leucine-rich repeat (NBS-LRR) gene families are considered to be critical components of plant immunity responses, which are categorized into TNLs and CNLs subclasses containing either a Toll/interleukin-1 receptor (TIR) or coiled-coil (CC) domain at the N terminus (Van de Weyer et al., 2019e Weyer et al., 2019). The number of NLRs in A. edgeworthii (102 genes) was markedly lower than that in most of the other species investigated (Table S20). A notable reduction in the number of TNL and CNL subclass genes was detected in A. edgeworthii, which accounted for about one-fourth and one-eighth of soybean and common bean, respectively (Figures 3b,c).

A model for the aerial and subterranean flower differentiation

In plants, the ABCE model genes which belong to the MIKC-type MADS-box gene family, play a key role in the determination of floral organ identity in angiosperms (Ng and Yanofsky, 2001). To investigate the basis for this divergence among different morphs of flowers, we identified 50 MADS-box genes, including homologues of the genes for the ABCE model of floral organ identities: AP1/FUL and AGL6 (A-class for sepals and petals), AP3/PI (B-class for petals and stamen), AG (C-class for stamen and carpel), and SEP1 (E-class for interacting with ABC function proteins) in A. edgeworthii (Figure 4a, Tables S21–S23). Phylogenetic analysis revealed that the MADS-box proteins were classified into 12 clades, including the ABCE classes in A. edgeworthii consistent with the established clades in Arabidopsis, soybean and common bean. However, the ratio of gene copies in ABCE classes in A. edgeworthii (number of ABCE classes genes/total MIKC-type gene copy numbers) was not significantly different between Arabidopsis, soybean and common bean, suggesting that different morphs of flowers do not result from gene copy number variation. We then compared the gene expression between the aerial CH flowers and the subterranean CL flowers of A. edgeworthii and found 3 major gene expression clusters (cluster 1–3) (Figure 4b; Figure S9). Cluster 1 contained the A-class homologues (AP1a/b), the B-class homologues (PIa/b), and the E-class homologues (SEPa/b/c/d) that are mainly characterized by high expression in flowers and seeds. The expression of all genes in cluster 1, except SEPd, was higher in the aerial flowers as compared to that in the subterranean flowers. Cluster 2 contained other A-class homologues (FULa/b/c), the SOC1 homologues (SOC1a/b/c), the SVP homologues (SVPa/b/c), and AGL12, which were highly expressed in the vegetative tissues such as leaf, stem, and roots. However, their expression was lower in the flowers and seeds. The 2 types A-class homologues (AP1s and FULs) had different expression patterns, suggesting that they might have undergone subfunctionalization for flower development after the legume WGD event. Contrary to the expression profiles of cluster 1, the SOC1 homologues (AGA1/2c and STK) in cluster 2, had significantly higher expression in subterranean flowers. Interestingly, an AP3 homologue (AP3, Ae00020092) with a predominant expression in aerial flowers is also under positive selection in A. edgeworthii (Table S24), which implies that the gene plays a key role in the differentiation regulation of aerial and subterranean flowers. We thus proposed a putative model for aerial and subterranean flower development in A. edgeworthii. The specifically morphological characters of petals, stamens, and ovules of the subterranean flowers may be attributed to the depressed expression of A/B/E-class genes and the enhanced expression of isoforms in C-class during the development process (Figure 4c).

Gene families involved in hard-seededness

Hard-seededness is a mechanism for maintaining seed dormancy and viability for long periods and is considered essential for the long-term survival of wild species (Sun et al., 2015). Seed-coat impermeability and hard-seededness are apparently different between aerial and subterranean seeds (Schnee and Waller, 1986). We investigated the seed coat anatomy by cross-section staining. From the outside to inside, the mature aerial seed coat is composed of cuticle, palisade cell layer with a light line on the outer edge close under the cuticle, hourglass cell layer, and parenchyma cells (Figure 5a). The seed coat of subterranean seeds is composed of a ‘pre-palisade’ cell layer and several layers of flattened parenchyma cells. Apparently, the cuticle and the light line is absent in the subterranean seeds. Coincidentally, the expansion of gene family in KEGG analysis suggests that the most significant enrichment pathways are involved in cutin, suberine, and wax biosynthesis (Table S16). Therefore, we next inspected the gene expression profiles of all genes known to be involved in cutin and wax biosynthesis. Generally, the majority of compounds comprising the cuticular wax are derived from very-long-chain (VLC) fatty acids (C20-C34), including alkanes, aldehydes, primary and secondary alcohols, ketones, and esters (Yeats and Rose, 2013). Two long-chain acyl-coenzyme A synthase (LACS) 1 and 3 LACS2 genes were identified in A. edgeworthii genome, and all of them had a higher expression in aerial seeds compared to that in subterranean seeds. The expression of KER1, PAS2, and CER10, which are the key genes for the fatty acid elongase (FAE) complex, was expressed at high levels in both aerial and subterranean seeds, whereas 6 CER6 genes showed a predominant expression in the aerial seeds. The transcriptomic data indicate that CER2, CER3, CER4, and WSD1 are highly expressed in the aerial seed but expressed at low levels in the subterranean seeds. However, midchain alkane hydroxylase (MAH1), which catalyses the production of secondary alcohols and their subsequent oxidation in ketones (Greer et al., 2007), presented with a very low expression in both the aerial and subterranean seeds, suggesting that VLC ketone may not be necessary for seed cuticle synthesis in A. edgeworthii. The cuticle biosynthesis genes CYP86A4 and GPA76 showed a relatively higher expression in
aerial seeds in comparison to that in the subterranean seeds. Homologues of CYP77A6, the CYP77 subfamily that encodes midchain hydroxylase, showed no differential expression differences between the aerial and subterranean seeds. Five homologues of ABCG32 and 2 of ABCG11 were identified in A. edgeworthii, demonstrating a prominent expression in aerial seeds (Figure 5b; Figure S10; Table S25). These genes are required for cutin accumulation and deposition in Arabidopsis (Li-Beisson et al., 2009).

**Discussion**

The genome size varied from 329 to 4275 Mb in the sequenced leguminous species, among of which A. edgeworthii possesses...
the smallest genome. It is well-documented that, in addition to polyploidization, retrotransposon amplification has been a major cause of genome expansion (Devos et al., 2002). Indeed, genome size is enormously variable and TEs, especially LTR retrotransposons, contribute greatly to the large variation observed among plant species. However, the key factors responsible for triggering LTR amplification remains unclear (Bennett and Leitch, 2005; Lee and Kim, 2014). Consistently, we found that the genome size is positively correlated with total TE length (specifically for the Typ3-gypsy subgroup) in assembled legumes with genome size varied from 0.30 Gb (A. edgeworthii) to 3.92 Gb (P. sativum). Accordingly, we broadly examined the synteny blocks between the genomes of A. edgeworthii and common bean, which are closely related species that underwent the same WGD events, to identify the contracted sequences in A. edgeworthii genome. However, we did not find any large ‘uncovered’ regions in common bean chromosomes, although we observed relatively more solo-LTRs in A. edgeworthii. Such LTRs are produced from unequal intrachromosomal homologous recombination, which indicates that recombination between long terminal repeats of LTR retrotransposons contributes to the compact size of the A. edgeworthii genome.

Amphicarpy is a rare phenomenon in angiosperms having been reported in only 100 species distributed across 13 phylogenetically distantly families, out of the nearly 250 000 flowering plant species (Kaul et al., 2000). Amphicarpy enables plants to exhibit life-history plasticity. Subterranean fruits allow viability and germination in fluctuating environments and avoid being grazed by most animals (Kaul et al., 2000; Lev-Yadun, 2000). Amphicarpic species invest more reproductive resources in subterranean flowers derived from the CL flowers than that in aerial flowers derived from the CH flowers under stressful conditions (Cheplick, 1987). Accordingly, the CL flowers are already simplified, including a partial suppression of petals and stamens, which could be a strategy to save energy under stressful conditions, although CL flowers scarcely manifest differences in the development process and morphology in comparison to those of aerial CH flowers (Zhang et al., 2006). Potential hybrids of aerial CH flowers may compensate for the genetic inbreeding depression associated with subterranean SCL reproduction (Culley and Klooster, 2007; Kaul et al., 2000). In CL species, CL flowers and fruits are considered as an adaptation to uncertain environmental conditions (Schnee and Waller, 1986).
Lipids are critical components of plant cell membranes and play important roles in many aspects. Lipids can provide energy for metabolic activities and serve a physical barrier on the surface of the epidermal cells to protect the plants from environmental assaults, and serve as second messengers in signal transduction involved in plant growth, development, and response to stresses (Shah, 2005). The cutin, suberin, and wax biosynthesis, and fatty acid biosynthesis pathway are among the mostly enriched pathways related to the lipid metabolism (Table S16). And the tissue-specific expression lipid synthesis and metabolism genes were observed in aerial subterranean seeds includes long-chain acyl-CoA synthase, midchain alkane hydroxylase, midchain hydroxylase, cuticle biosynthesis (Figure S5b, Table S25). These data suggested that lipid synthesis and metabolism may involve in the regulation of cutin and wax biosynthesis that contribute to differentiate the seed-coat impermeability and hard-seededness of aerial and subterranean seeds in A. edgeworthii.

Serine/threonine protein kinases are involved in the response to environmental stresses (such as high salt, drought, and low temperature) that regulate multiple life processes in cells. It has been speculated that the contraction of these types of gene families affect the accumulation of secondary metabolites and reduces environmental adaptability, resulting in the survival of Cladopus chinensis only in good-quality water (Xue et al., 2020). Similarly, A. edgeworthii is sensitive to various environmental stresses and is difficult to cultivate in the greenhouse. We thus speculated that the contraction of the above-mentioned gene families in A. edgeworthii has been a major driving force in the evolution of the survival strategy amphicarpy. In other word, amphicarpy is a bet-hedging strategy in response to biotic and abiotic stressors (Zhang et al., 2020). Although amphicarpy has evolved independently in phylogenetically unrelated taxa; however, the extent of molecular convergence involved remains unknown. For instance, the genetic analysis showed the non-amphicarpy is dominant to the amphicarpy by cross between the four perennial amphicarpic morphotypes and a non-amphicarpic species would favour to uncover the mechanisms behind this enigmatic life-history strategy.

Amphicarpic plants can produce both aerial seeds and subterranean seeds, and the subterranean fruits of these plants are formed earlier than the aerial ones and the amount of the aerial fruits were determined by their growth conditions (Cheplick, 1987; Zhang et al., 2020). According to the specific phenomenon, the idea of the optimistic strategy and pessimistic strategy for plant fruit setting was firstly put forward (Zeide, 1978). It seems that there exists a rheostat to balance the allocation to aerial fruits and subterranean ones according the environments (Cheplick and Quinn, 1982; Kaul et al., 2002; Kawano et al., 1990; Weiss, 1980). For instance, the proportion of CL seeds increased and SCL seeds increased accordingly with the fungal infection intensity enhancement (Parker, 1986). This mechanism could guarantee enough progenies for next generation even in very extreme conditions and they do not need high stress tolerant ability for the amphicarpic species, which could be the cause that the stress-related gene family contracted (Figure 3). All the phenomena evolved with the genome variation during evolution, however, the genome could not explain everything currently. But we believe that the genome of A. edgeworthii will provide a basis for fruit amphicarpy and the distinct morphology of flowers including chasmogamous and cleistogamous evolution. As it possesses the smaller genome size and abundant seeds can produce in three to four months, we suggested that A. edgeworthii can be developed as a model species for the studies on amphicarpy and distinct morphology evolution of flowers.

Methods

Sampling and sequencing

The A. edgeworthii plants used in this study grew in the Qianfo Mountain (Jinan, China. N36°37′40.59″, E117°02′49.81″). The leaves were collected for massively parallel sequencing (MPS) and nanopore sequencing. The tissue samples were collected for transcriptomic analysis including aerial flowers, subterranean flowers, leaves, pod shells, roots, stems, aerial seeds, and subterranean seeds, respectively from seedling grown in greenhouse as previously described (Zhang et al., 2006). The whole aerial flowers and subterranean flowers were collected at middle age. We collected the whole unripe aerial seeds in the green pods (approximately 35 days) and subterranean seeds in the similar stage and seed coats from aerial and subterranean seeds that were easily to peel. All samples were immediately frozen by liquid nitrogen and stored at −80°C until DNA/RNA extraction.

For MPS sequencing, 10 μg of genomic DNA was sheared by ultrason and fragments with 200–400 bp in length were selected for DNA sequencing performed using the DNSEQ platform (BGI, Shenzhen, China). The BGISEQ-500 sequencer was used to prepare the library and produce the sequencing data. SOAPnuke (v1.5.6) was used to filter reads with the parameters ‘-n 0.01 -l 15 -q 0.3 -i -Q 2 -G -M 2 -A 0.5 -d’. Low-quality reads that contained adaptors and reads with ambiguous base ratio higher than 1% or ratio of low-quality bases (quality score < 20) higher than 10% were removed (Chen et al., 2018).

For Oxford nanopore sequencing, the sequencing libraries were constructed following the protocol for Oxford nanopore’s ‘1D gDNA selection for long reads’. For each library, 10 μg of high molecular weight DNA was sheared by ultrasound, and the fragments were selected using the BluePippin. The selected DNA fragments were end repaired, followed by adapter ligation, and the final library was eluted in the elution buffer. The libraries were loaded onto primed R9.4 Spot-On Flow cells for sequencing on the GridION × five sequencers. The Oxford Nanopore Albacore software (v2.1.3) was used to perform base-calling analysis of the unprocessed data (Schmidt et al., 2017).

De novo genome assembly

The high-quality MPS data of 77.12-Gb of 150-bp paired-end reads were generated on the DNSEQ platform, and the genome size and heterozygosity were estimated using a k-mer size of 17 bp by the Jellyfish (v2.2.4) and GenomeScope (Marciais and Kingsford, 2011; Vurture et al., 2017). The genome was assembled as specified the previous study (Schmidt et al., 2017).

Canu (v1.8) was used to correct errors with the parameters: -corrOutCoverage 500 -corrMinCoverage 2 -minReadLength 2000. The corrected longest reads with 30× coverage were assembled into contigs using the SMARTdenovo with k = 17. After finishing the initial assembly, Racon and Medaka were adopted to perform the further correction with default parameters recommended. Finally, Pilon (v1.22) was used to further polish the assembled genome using default parameters.

Hi-C library construction and sequencing were performed as previously described (Dudchenko et al., 2017; Durand et al.,...
Briefly, freshly harvested leaves were fixed in 2% formaldehyde, and the blunt-end DNA strands were digested with the HindIII enzyme and used for Hi-C library construction by using the Hi-C Plant Kit. The data of 150-bp paired-end reads generated on the DNSEQ platform were mapped to the assembled contigs using Bowtie. The mapped datasets were submitted to the Juicer and 3d-dna pipeline for grouping, ordering, and orientation of the contigs. The Benchmarking Universal Single-Copy Orthologs (BUCSO) was used to assess the quality of assembled genome (Simao et al., 2015). Transcriptome sequencing and analysis

The total RNA of tissues from A. edgeworthii, including leaves, aerial flowers, aerial seeds, subterranean flowers, subterranean seeds, and roots, were extracted with Trizol as described previously (Cui et al., 2018). The quality of RNA was detected by 1.5% agarose gel electrophoresis, and the concentration was detected by the NanoDrop™ One UV-Vis Spectrophotometer (Thermo Fisher Scientific, USA). The paired-end sequencing with 150-bp was conducted on the BGISEQ-500 sequencer (BGI, Shenzhen, China). The clean data was generated by removing reads containing poly-N, adapters or low-quality reads from the raw data. Mixed RNA from all samples was sequenced to generate de novo transcriptome for genome annotation using Trinity (Haas et al., 2013). For gene expression quantification in tissues, fragments per kilobyte of exon model per million reads mapped (FPKM) values were calculated using the HISAT2 (v2.1.0) and Cufflinks (v2.2.1) (Kim et al., 2015; Trapnell et al., 2010).

Gene annotation

Gene annotation was performed as previously described with minor modification (Schmidt et al., 2017). Marker (v2.31.8) was used to perform the first-round of homologous gene annotation based on protein sequences from 6 legumes including Cicer arietinum, Glycine max, Lupinus angustifolius, Medicago truncatula, Phaseolus vulgaris, and Vigna angularis. The Augustus and SNAP were employed for de novo annotation. RNA-seq data was used for annotation by using the HISAT2 and StringTie (Kim et al., 2015). Finally, Marker was used to conduct the second-round of annotation based on the predicted transcripts and the August and SNAP models (Holt and Yandell, 2011). TEs in the genome were predicted by combining homology-based and de novo strategies. The RepeatMasker, RepeatProteinMask, and LTR_FINDER were employed for homology-based prediction and de novo identification (Tarailo-Graovac and Chen, 2009; Xu and Wang, 2007). Tandem repeats were predicted using Tandem Repeats Finder (v4.0.9/Benson, 1999). Gene functional annotation was performed using BLAST (v2.2.31) (Altschul et al., 1990). The annotated genes were aligned against SwissProt (http://www.uniprot.org/), TrEMBL (http://www.uniprot.org/), Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/), InterPro (https://www.ebi.ac.uk/interpro), and KOG and GO. tRNA, miRNA and snRNA, and rRNAs were annotated by tRNAscan-SE (v1.3.1), INFERNAL, and BLASTN, respectively (Lowe and Eddy, 1997). Transcription factor genes in the genomes were identified by comparing with corresponding genes in other plant genomes using the iTAK and HMMscan (Potter et al., 2018; Zheng et al., 2016).

Genome evolution

The single-copy gene families were used to construct phylogenetic tree by using the PhyML (v3.0), and V. vinifera was designated as an outgroup (Guindon et al., 2009). The divergence time of the nodes on the phylogenetic tree was referred by the MCMCTree in PAML (v4.5) (Yang, 1997). The following calibration times of divergence were obtained from the TimeTree database (http://www.timetree.org/): 107-135 Mya for V. vinifera and L. angustifolius, 30-54 Mya for M. truncatula and C. arietinum, 53-85 Mya for L. angustifolius and M. truncatula, 11.7-27.5 Mya for G. max and C. cajan.

BLASTP (E-value < 1e-10) was referred to query the homologues and MCScanX were used to identify syntenic blocks within the genome of A. edgeworthii and plotted with Circos (Krzywinski et al., 2009; Tang et al., 2008). Ks values were plotted to identify the putative whole-genome duplication events by analysing of gene pairs in the synteny blocks using PAML (v4.5) (Yang, 1997). The 4DTv distances of syntenic blocks were calculated and corrected using the HKY model (Hasegawa et al., 1985; Huang et al., 2009). Protein sequences of V. vinifera, G. max, M. truncatula, C. arietinum, P. vulgaris, L. angustifolius, V. angularis, G. soja, and C. cajan were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). The identification of gene families was performed by using BLASTP (v2.2.26) with an e-value threshold of 1e-10 based on sequence alignments (Huang et al., 2020). CAFE was used to identify the expanded and contracted gene families between A. edgeworthii and the other 9 species mentioned above (De Bie et al., 2006). The expanded and contracted gene families were further subjected GO analysis (Ashburner et al., 2000).

The protein-coding sequences of each single-copy gene family were aligned using prank (v.100802) with an empirical codon model, and the codon-based alignments were further filtered using Gblocks (v0.91) to trim the poor-aligned regions. The Ka/Ks ratios (the number of nonsynonymous substitutions per nonsynonymous site, Ka, over the number of synonymous substitutions per synonymous site, Ks) were calculated using the CodeML tool in the PAML package (v4.9) with the branch site model.

Resistance-related genes that were related to domains such as the LRR and TIR were detected by pHMM search using the HMMER (v3.0) software package (Potter et al., 2018), and the prediction of the coiled-coil domains was performed using Paircoil2 (McDonnell et al., 2006). We further aligned the protein sequences using MAFFT and build phylogenetic trees using FastTree separately for both the CNL and TNL types of R genes.

Solo-LTR and intact LTR prediction

The candidate LTR retrotransposons (LTR-RTs) elements were identified using the LTR_finder (v1.07) with a default setting, and LTRharvest tool with ‘-maxlenLtr 7000 -maxsd 6 -vic 10’ parameters as well as 1-bp mismatch was allowed to start and end the motif ‘TGCA’ (Xu and Wang, 2007). LTR_retriever (v2.8) program was implemented for the accurate identification of LTR retrotransposons from outputs of LTR_finder and LTRharvest (Ou and Jiang, 2018). After filtering out the tandem repeats and the elements with gaps and those of unusual size, sequence elements were classified based on the order of their protein domains in the internal region. Non-redundant high-quality libraries were generated following the internal boundaries detected, TSD and motif identification, non-LTR proteins, and nested insertion cleaning. Then, whole-genome LTR-RTs were annotated based on the libraries. Furthermore, the solo LTRs were detected from the whole-genome annotation of LTR elements with the Perl script ‘solo_finder.pl’ in the LTR_retriever software. Intact LTR-RTs with a pair of LTRs span of 85–5000 bp at both the termini as well as...
in the internal region including the protein domains, were also identified using the 'intact_finder_coarse.pl' script in the LTR_retriever. The ratios of solo LTRs to intact LTRs in each of the LTR family were calculated using the script 'solo_intact_ratio.pl'.

The whole genomes were scanned for possible TE domains using the DANTE (RepeatExplorer server; https://repeatexplorer-elixir.cerit-sc.cz/) software. In this process, the database of Viridiplantae protein domains derived from TEs domain were searched using lastal (v1.047) (Frith et al., 2010). Matches with at least 80% coverage, 35% identity, and 45% similarity were considered as homologous and annotated. The phylogenetic trees were separately constructed for copia and gypsy LTR using reverse transcriptase (RT) domain for representation. The protein sequences of RT domains of these 2 types of LTR from each species were extracted and merged together. Then the multiple alignment was generated for each TE type using the MAFFT (v7.453) and further feed into FastTree (v2.1.10) for tree building. To further estimate the expand time of these two types of LTR in each species, the protein sequences of gypsy RT domain were used as an outgroup for copia repeat tree, while the sequences from copia RT domain were used as an outgroup for gypsy. The distance of each terminal node to the root node was calculated using the ete3 Python package (Huerta-Cepas et al., 2016).

For the estimation of LTR insert time, the proportion of sequence differences between the 5’ and 3’ direct repeats of an LTR candidate from the LTR_retriever output was approximated as 100% - identity%. The sequence identity was estimated using the BLAST (v2.7.1). The divergence time, which provides the time of insertion, was calculated as \( T = K/2q \), where \( K \) was estimated under the Jukes-Cantor model using the divergence value, and the neutral mutation rate was set as \( 1.3 \times 10^{-8} \) mutations per site per year.

Real-time Quantitative RT-PCR (q-RT-PCR)

The transcript abundance was investigated as described (Cui et al., 2018). Briefly, total RNA was extracted as described as sample preparation for RNA-seq. The cDNA was synthesized using PrimeScriptTM RT reagent Kit with gDNA Eraser (TaKaRa, RR047A). Eight genes were randomly selected to verify the RNA-sample preparation for RNA-seq. The cDNA was synthesized.

Acknowledgements

This work was supported by the Taishan Scholar Program [No. tsqn20161058], National Natural Science Foundation of China [31871665] and the Innovation Program of SAAS [CXGC2019G02]. All the authors appreciate Prof. Weiliang Ma friendly provided partial morphology images of A. edgeworthii (see Figures 1 and 4).

Author contributions

W.S., L.G., and L.Y. conceived the project and designed the study. Z.X., X.G., L.G., and L.Y. performed the sampling and experiments. L.Y., H.K., F.G., and L.G. performed the data analysis. L.Y. designed and visualized the figures. F.G., L.Y., L.G., S.I., and W.S. wrote the manuscript. All authors read and approved to publish the final manuscript.

Conflict of interests

The authors declare no competing conflict interests.

Data Availability Statement

DNA sequencing data are available in BioProject of CNSA (https://db.cngb.org/cnsa/) under accession number CNP0001124. The RNA-seq data are available under the accession number CNX0196691. The final genome assembly is available under the accession number CNA0014149. And also, all the data deposited in BioProject of GenBank under accession numbers PRJNA644665.

References

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

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P. et al. (2000) Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.

Ballard, H.E., Sytsma, K.J. and Kowal, R.R. (1998) Shrinking the violets: phylogenetic relationships of infrageneric groups in Viola (Violaceae) based on internal transcribed spacer DNA sequences. Syst. Bot. 23, 439–458.

Barker, N.P. (2005) A review and survey of Basicarpy, Geocarpy, and Amphicarpy in the African and Madagascan Flora. Ann. Mo. Bot. Gard. 92, 445–462.

Bennett, M.D. and Leitch, I.J. (2005) Genome size evolution in plants. Evol. Genome, 89–162.

Bennett, M.D. and Leitch, I.J. (2012). Plant DNA C-values database (release 6.0, Dec. 2012).

Bennetzen, J.L., Ma, J. and Devos, K.M. (2005) Mechanisms of recent genome size variation in flowering plants. Ann. Bot. 95, 127–132.

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

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y. et al. (2018) SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience 7, 1–6.

Cheplick, G.P. (1987) The ecology of amphicarpy plants. Trends Ecol. Evol. 2, 97–101.

Cheplick, G.P. and Quinn, J.A. (1982) Amphicarpum purshii and the “pessimistic strategy” in amphicarpic annuals with subterranean fruit. Oecologia, 52, 327–332.

Cui, F., Su, N., Duan, G., Liu, Y., Han, Y., Liu, S., Wan, S. et al. (2018) Identification of metabolites and transcripts involved in salt stress and recovery in peanut. Front. Plant Sci. 9, 217.

Culley, T.M. and Klooster, M.R. (2007) The cleistogamous breeding system: a review of its frequency, evolution, and ecology in angiosperms. Bot. Rev. 73, 1–30.

De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics, 22, 1269–1271.

Demuth, J.P. and Hahn, M.W. (2009) The life and death of gene families. BioEssays, 31, 29–39.

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

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S. et al. (2017) De novo assembly of the Aedes aegypti Genome Res.
status and genomic characteristics of Cladopus chinensis (Podostemaceae).  
Hortic. Res. 7, 1–16.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood.  
Comput. Appl. Biosci. 13, 555–556.

Yeats, T.H. and Rose, J.K. (2013) The formation and function of plant cuticles.  
Plant Physiol. 163, 5–20.

Zeide, B. (1978) Reproductive behavior of plants in time.  
Am. Nat. 112, 636–639.

Zhang, K., Baskin, J.M., Baskin, C.C., Yang, X. and Huang, Z. (2015) Lack of divergence in seed ecology of two Amphicarpaea (Fabaceae) species disjunct between eastern Asia and eastern North America.  
Am. J. Bot. 102, 860–869.

Zhang, Y., Yang, J. and Rao, G. (2005) Genetic diversity of an amphicarpic species, Amphicarpaea edgeworthii Benth. (Leguminosae) based on RAPD markers.  
Biochem. Syst. Ecol. 33, 1246–1257.

Zhang, Y., Yang, J. and Rao, G. (2006) Comparative study on the aerial and subterranean flower development in Amphicarpaea edgeworthii Benth. (Leguminosae: Papilionoideae), an amphicarpic species.  
Int. J. Plant Sci. 167, 943–949.

Zheng, Y., Zuo, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P., Barf, M. et al. (2016) iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases.  
Mol. Plant, 9, 1667–1670.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Genome survey of A. edgeworthii using 17-mer analysis.

Figure S2 The diagram for length distribution of clean data.

Figure S3 Hi-C clustering for pseudomolecule construction in A. edgeworthii.

Figure S4 The integration high-quality genome A. edgeworthii of genetic data.

Figure S5 Comparisons of the distribution of gene, coding sequences (CDS), exon and intron length for protein-coding genes in the genomes of A. edgeworthii and other leguminous species.

Figure S6 Genome synteny between A. edgeworthii and common bean, soybean.

Figure S7 Genome synteny between A. edgeworthii, common bean, adzuki bean, and soybean chromosomes.

Figure S8 DNA C-values of A. edgeworthii and other leguminosae species.

Figure S9 Real-time quantitative PCR analysis of selected genes in Flower_Aerial and Flower_Sub.

Figure S10 Real-time quantitative PCR analysis of selected genes in Seed_Aerial and Seed_Sub

Table S1 Global statistics of the Nanopore sequencing data for the A. edgeworthii genome.

Table S2 Statistics of the assembled contigs.

Table S3 Statistics of Hi-C library reads by Juicer software.

Table S4 Statistics of genome sequence anchored onto pseudo-molecules.

Table S5 Statistics of the assembled scaffold using 3D-DNA.

Table S6 Statistics of gene prediction results based on three methods.

Table S7 Functional annotation of protein-coding genes for A. edgeworthii.

Table S8 Statistics of non-coding RNAs in the A. edgeworthii genome.

Table S9 The quality evaluation of Hi-C assemblies genome by BUSCO software.

Table S10 The quality evaluation of gene prediction by BUSCO software.

Table S11 Genome size and gene number.

Table S12 The correlation between genome size and the size of different TE classes.

Table S13 Comparison of Ty1-copia and Ty3-gypsy copy numbers in A. edgeworthii and other species.

Table S14 Comparison of solo-intact (S/I) ratio of LTR retrotransposons in three plant genomes.

Table S15 KEGG pathway enrichment of A. edgeworthii specific gene families.

Table S16 The KEGG pathway enrichment of A. edgeworthii significantly expanded gene families.

Table S17 The GO pathway enrichment of A. edgeworthii significantly expanded gene families.

Table S18 The KEGG Pathway enrichment of A. edgeworthii significantly contracted gene families.

Table S19 The GO Pathway enrichment of A. edgeworthii significantly contracted gene families.

Table S20 Comparison of NBS-LRR gene copy numbers in A. edgeworthii and other species.

Table S21 Information of MADS genes in Amphicarpaea edgeworthii.

Table S22 Information of MIKC⁺ genes in Amphicarpaea edgeworthii, Arabidopsis, Glycine max, Phaseolus vulgaris.

Table S23 The FPKM of MIKC⁺ genes in Amphicarpaea edgeworthii.

Table S24 Positive selection genes and KEGG enrichment in Amphicarpaea edgeworthii.

Table S25 Cutin and Wax biosynthesis pathway genes and FPKM value of gene expressions.

Table S26 Primers used in this study for qRT-PCR.

© 2020 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 19, 952–965