Identification and characterization of abundant repetitive sequences in *Allium cepa*

Jiaping Fu1,4, Hao Zhang2,4, Fengling Guo1, Lu Ma3, Jinping Wu1, Mengxia Yue2, Xueke Zheng2, Zhengming Qiu1* & Lijia Li2*

Species of the genus *Allium* are well known for their large genomes. *Allium cepa* is of great economic significance. Among vegetables, it ranks second after tomato in terms of the global production value. However, there is limited genomics information available on *A. cepa*. In this study, we sequenced the *A. cepa* genome at low-coverage and annotated repetitive sequences by using a combination of next-generation sequencing (NGS) and bioinformatics tools. Nearly 92% of 16 Gb haploid onion genome were defined as repetitive sequences, organized in 162 clusters of at least 0.01 percent of the genome. Of these, a proportion representing 40.5% of the genome were further analyzed in detail to obtain an overview of representative repetitive elements present in the *A. cepa* genome. Few representative satellite repeats were studied by fluorescence in situ hybridization (FISH) and southern blotting. These results provided a basis for evolutionary cytogenomics within the *Allium* genus.

The C value, which is the DNA content in a gamete, may vary widely among closely related organisms. Large comparative analyses of the plant DNA C-value database (http://data.kew.org/cvalues/) have shown great variation in terms of genome size in plants. Part of this variation is due to differences in the ploidy level, while the majority of the variation is based on differences in the repeats. There is a lack of correlation between genome size and complexity of eukaryotic genomes, which is called C-value paradox. The main reason for genome size variations are differences in the abundance of transposable elements (TEs), while gene numbers remain relatively constant.

The genus *Allium* (Amaryllidaceae) belongs to monocotyledonous genera and includes more than 800 species, which are widely distributed over the Holarctic region from the dry subtropics to the boreal zone. This genus includes the most common chromosome number x = 8 and other numbers (e.g., x = 7, 9, 10, 11) and a number of polyploids are characterized by a large genome size. The genome size of 30 *Allium* species varies from 7 pg (*A. allynolicum*, 2n = 4 x = 32) to 31.5 pg (*A. ursinum*, 2n = 2 x = 14). The genus *Allium* is of great economic importance because it includes several important vegetable crops and ornamental species, such as *A. cepa* L. (onion), *A. sativum* (garlic), *A. schoenoprasum* (chives), *A. wakegi* (scallion), *A. cepa* var. aggregatum (shallot), and *A. ampeloprasum* var. porrum (leek). In terms of the global production value, onion ranks second after tomato. While the tomato genome is fully sequenced, little is known about the onion genome. This is partly attributed to the giant size of the onion genome (1 C = 16 Gbp), which consists of more than 95% of repetitive elements.

An increasing number of plant genomes are fully sequenced (https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes). But it is still a challenge and costly to sequence large genomes such as that of *A. cepa*. Therefore, in the onion genome sequencing project (SEQUON, http://www.oniongenome.net), it is proposed to only sequence gene-rich regions combining transcriptome sequencing and enrichment of low-copy regions by removing repetitive DNAs. Until now, only reference gene sets have been revealed by transcriptome sequencing. The study of repetitive sequences in onion has mainly focused on tandem repeats, including the 375 bp satellite sequence family and telomeric repeats. Mancia *et al.* (2015) analyzed the distribution of rDNA, cot-1 DNA and 375 bp satellite in two *A. cepa* varieties by FISH. There was no comprehensive analysis of repetitive sequences in *A. cepa*. Until recently, a comparative study of the repeatome on three *Allium* species, including onion has been performed.

1Institute of Economic Crops, Hubei Academy of Agriculture Sciences, Wuhan, 430064, China. 2State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan, 430072, China. 3Shenzhen Tobeacon Technology Co. Ltd., Shenzhen, 518049, China. 4These authors contributed equally: Jiaping Fu and Hao Zhang. *email: 13808640602@163.com; ljli@whu.edu.cn*
dehydrated with an ethylalcohol series. After that the root tips were digested by 2% cellulase and 2% photolyase for 60 min at 37 °C. The digested DNA transposons transpose through a DNA intermediate via a “cut-and-paste” mechanism, usually maintaining a moderate copy number in the genome. In this study, we sequenced the A. cepa genome at low-coverage, identified and characterized its most abundant repetitive sequences and determined the chromosomal localization of a few repeats.

Materials and Methods

Materials. A. cepa cv. Fuxing was obtained from Institute of Economic Crops, Hubei Academy of Agriculture Sciences, China.

DNA extraction for NGS. We collected leaves and roots from the greenhouse grown seedlings and carried out DNA extraction using a DNeasy plant mini kit from Qiagen.

NGS. A sequencing library was prepared using NEB Next™ Ultra™ DNA Library Prep Kit Illumina (New England, Biolabs, Ipswich, MA, USA). Paired-end sequencing (2 × 150 bp, 350–400 bp insert size) of total genomic DNA was performed by iGeneTech Co. Ltd. (Beijing, China) on the Illumina HiSeq2500 platform on a single lane. Clean sequencing data were supplied in FASTQ format without adapters.

Repeat explorer. The RepeatExplorer pipeline (https://repeatexplorer-elixir.cerit-sc.cz/galaxy/) was performed to cluster NGS raw reads into groups of similar reads with default setting. Repeat clusters with genome proportions of no less than 0.01% were selected for further analysis. Repeat clusters with known protein domains can be classified by RepeatExplorer pipeline directly. Other clusters were subjected to manual analysis with similarity searches against GenBank databases (Nt and Nr) using Blastn and Blastx with an E-value of 1e-5.

PCR amplification, cloning, and sequencing of AceSat01–377 and AceSat02–750. Amplification of the determined repeats was performed by using specific primers (Table 1) designed by NCBI Primer blast (https://www.ncbi.nlm.nih.gov/tools/primer-blast). The PCR conditions were 98 °C 1 min, 35 cycles: 98 °C 15 s; 60 °C 15 s; 72 °C 30 s; final elongation: 72 °C 3 min. The sequences of the repeats were validated by cloning of the PCR product into pGEM-TEasy vector (Promega, Madison, WI, USA) according to manufacturer’s instruction. The individual clones were sequenced using an ABI 3130xl Genetic Analyzer. The sequences of the repeat units were submitted to NCBI GenBank (Table 1).

Chromosome preparation. The chromosomes are prepared as mentioned by He et al. Root tips of A. cepa cv. Fuxing were collected when they reached 0.5–1 cm. Mitosis was blocked in α-bromonaphthalene avoiding light at room temperature for 4 hours, followed by fixing in 4% (w/v) paraformaldehyde, stored at 4 °C for 40 min. After that the root tips were digested by 2% cellulase and 2% photolyase for 60 min at 37 °C. The digested root tips were homogenized in 60% acetic acid solution and dropped on glass slides. The prepared slides were dehydrated with an ethylalcohol series.

Fluorescence in situ hybridization (FISH). pTa71 of 45 S rDNA was used in the present investigation. The repeated DNA sequences used for FISH were labelled by nick translation with RED: Texas Red-12-dUTP (Invitrogen C3176) and GREEN: Fluor 488–5-dUTP (Invitrogen C11397) FISH was performed as described. The slides were treated in 70% formamide in 2 × saline sodium citrate (SSC) for 5 min at 90 °C. Simultaneously, 2 mg/ml probes and 1 mg/ml sheared salmon sperm DNA were pre-mixed in 2 × SSC with 10% dextran sulphate and 50% deionized formamide, denatured at 80 °C for 5 min and immediately cooled in ice water. After dehydrating and air drying, the slides were incubated in denatured hybridization solution overnight at 37 °C. Nuclei and chromosomes were stained with 4, 6- diamidino-2-phenylindole (DAPI, 0.2 mg/ml, Sigma, Deisenhofen, Germany) and observed under an Olympus BX-60 fluorescence microscope. Images obtained using a CCD monochrome camera Sensys 1401E were pseudo-colored and processed with the Metamorph imaging system (Universal Imaging Corp., PA, USA. version 4.6.3) and Adobe Photoshop 9.0 software.

| Tandem repeats | Primers 5′-3′ | Accession number in NCBI | Size of repeat unit (bp) | Expected length of PCR product (bp) |
|----------------|--------------|--------------------------|-------------------------|-----------------------------------|
| AceSat02–750   | F:TCACACATgTgACACTgCgATATTAAAAg | MH017541 | 750 | 802 |
|                | R:TTTATCCgTCCgTGgTgATCCA     |         |     |      |
| AceSat01–377   | F:gATATgTgCATCATCgCAGAgAgg | MH017542 | 377 | 610 |
|                | R:ggTGcGCgggAAAAAAATgAagg   |         |     |      |

Table 1. Summary of the tandem repeats.

Based on advances in next-generation sequencing (NGS), de novo assembly and annotation methods, repetitive sequences can be studied effectively at reasonable costs by combining low-pass NGS and graph-based clustering analysis. Differences in genome sizes between Arabidopsis (1 C = 150 Mb) and A. cepa (1 C = 16 Gb) are mainly due to the amplification of repetitive DNA sequences. Repetitive sequences include tandem repeats (satellites, minisatellites and microsatellites) and transposable elements (TEs). As in all other eukaryotes, TEs in plants are categorized as class I (retrotransposons) or class II (DNA transposons) transposons. Class I (retrotransposons) contains all TEs that transpose via an RNA intermediate in a “copy-and-paste” process. Class II DNA transposons transpose through a DNA intermediate via a “cut-and-paste” mechanism, usually maintaining a moderate copy number in the genome. In this study, we sequenced the A. cepa genome at low-coverage, identified and characterized its most abundant repetitive sequences and determined the chromosomal localization of a few repeats.
the genome, which are likely underestimated. Kiseleva de novo repeats as degenerative copies or if they are unrelated to them stays to be answered. Hertweck and Bainard used determined due to the RepeatExplorer threshold for clustering. The question if they come from the mainstream alkaline phosphatase (AP) conjugated anti-DIG antibody (1:5000) and chemiluminescence visualization with Starter Kit I (Cat.No.11 745 832 910, Roche Diagnostics GmbH, Germany). Detection was performed using the DIG High Primer DNA Labeling and Detection Mannheim, Germany). Hybridization was performed using the DIG High Primer DNA Labeling and Detection Starter Kit I (Cat.No.11 745 832 910, Roche Diagnostics GmbH, Germany). Detection was performed using alkaline phosphatase (AP) conjugated anti-DIG antibody (1:5000) and chemiluminescence visualization with 5-bromo-4-chloro-3-indolyl phosphate/ nitro blue tetrazolium (BCIP/NBT).

Results and Discussion

Genomic repeatome composition. Because of its large genome size of about 16 Gbp/1 C, it is difficult to analyze the repeat composition of the whole onion genome by traditional molecular methods. Thus, we used the latest NGS technology and RepeatExplorer computational pipeline to reveal the genome structure. A total of 15,990,607 clean paired-end reads were obtained with a length of 150 bp each. Illumina sequencing can lead to a bias at the beginning and the end of reads. Therefore, we trimmed the 10 bp on both ends of each read. Reads with quality score \( Q > 10 \) over 95% of bases without Ns were further analyzed. Only complete read pairs (interlaced reads) were used. The RepeatExplorer pipeline revealed us to use 0.01–0.5 \% genome coverage reads for analysis (https://repeatexplorer-elixir.cerit-sc.cz). To maintain a balance between high sensitivity and moderate running times with the available computational resources, we decided to use 2,642,364 reads representing ca. 2.16\% of the genome for clustering in RepeatExplorer pipeline. The RepeatExplorer pipeline arranged 2,419,056 reads in 537,598 clusters. The genomic composition of clusters generated by RepeatExplorer of 2,642,364 reads (2.16\% of genome coverage). And the RepeatExplorer pipeline arranged 2,419,056 reads in 537,598 clusters. The genomic proportion of these identified repeats was about 92%. X-axis: cumulative proportion of the clusters of the genome. Y-axis: numbers of reads.

Southern blotting. Genomic DNA was digested by restriction enzymes EcoRI and XbaI respectively at 37 °C for 12 h. Southern blotting was carried out using the method described by Li et al. Nylon transfer membranes were MILLIPORE.IMMOBILON-NY + 30 cm × 3.3 m Roll (Catalog No. INYCO010). Repeated DNA sequences used for southern blotting were labelled by nick translation with digoxin-11-dUTP (Roche, Mannheim, Germany). Hybridization was performed using the DIG High Prime DNA Labeling and Detection Starter Kit I (Cat.No.11 745 832 910, Roche Diagnostics GmbH, Germany). Detection was performed using alkaline phosphatase (AP) conjugated anti-DIG antibody (1:5000) and chemiluminescence visualization with 5-bromo-4-chloro-3-indolyl phosphate/ nitro blue tetrazolium (BCIP/NBT).

Figure 1. Repeat composition of clusters generated by RepeatExplorer of 2,642,364 reads (2.16\% of genome coverage). And the RepeatExplorer pipeline arranged 2,419,056 reads in 537,598 clusters. The genomic proportion of these identified repeats was about 92%. X-axis: cumulative proportion of the clusters of the genome. Y-axis: numbers of reads.
rather than personal errors between different laboratories (Fig. S6). However, Peška et al. (2019) used PCR free library for NGS, while we used normal library for sequencing. In addition, the difference might be attributed to intraspecific variability caused by different cultivars used in both studies.

PCR amplification, cloning and sequence analysis of satellite DNAs. Tandem repetitive DNA is an important component of the repetitive sequences. According to the length of the repeat arrays and their sizes, tandem repetitive DNA sequences can be classified into three groups: (i) microsatellites with 2–5bp repeats and an array size of the order of 10–100 units, (ii) minisatellites with 6–100 bp (usually around 15 bp) repeats and an array size of 0.5–30 kb and (iii) satellite DNA (satDNA) with a variable AT-rich repeat unit that often forms arrays up to 100 Mb, suitable as chromosome markers in FISH experiments8,28.

NGS and graph-based clustering analyses provide high-throughput tools for detecting satellite DNA28,29. As suggested by Ruiz-Ruano et al.29, the satDNA terminology should begin with species abbreviation in Repbase (e.g. Ace for Allium cepa) followed by the term “Sat”, a catalog number in order of decreasing abundance (according to the first genome analyzed), followed by consensus monomer length. Therefore, we termed two satellites as AceSat01–377 for cluster 7 (AcCL7) and AceSat02–750 for cluster 43 (AcCL43). Such satellite DNA usually showed a globula-like (AceSat01–377) or ring-like (AceSat02–750) graph30,31 (Fig. 2A). The AceSat01–377 and AceSat02–750 clusters comprised of 1.24% and 0.17% of the genome, respectively. The monomer length of satDNA sequences ranges from 150–400 bp in majority of plants and animals15. The length of the monomers for AceSat01–377 and AceSat02–750 are 377 bp and 750 bp, respectively (Table 1). These two repeats have been cloned and sequenced (GenBank accession numbers are MH017542 for AceSat01–377 and MH017541 for AceSat02–750). AceSat01–377 showed high similarity with a satellite commonly found in Allium species32 and low diversity among monomers. For AceSat02–750, BLASTn analysis revealed no significant match against National Center for Biotechnology Information (NCBI) databases (e–value = 10−5), suggesting that this might be a novel satellite repeat.

Table 2. Repeat composition of Allium cepa genome estimated from the Illumina sequencing data.

| Repeat type                  | Genome proportion (%) |
|------------------------------|-----------------------|
| Low_complexity               | 8.599                 |
| Simple_repeat                | 1.912                 |
| Satellite                    | 1.421                 |
| rDNA                         | 0.581                 |
| DNA.CMC.EnSpm                | 0.528                 |
| LINE.L1                      | 0.22                  |
| Plastid                      | 1.05                  |
| Unknown in analyzed clusters | 8.393                 |
| Total in analyzed clusters   | 40.5                  |
| Small clusters that were not analyzed | 5.049 |
| Non-clustered reads          | 8.451                 |
| Total                        | 100                   |

Figure 2. (A) The layout of AceSat01–377 and AceSat02–750 clusters from RepeatExplorer. (B) PCR amplification with primers for AceSat01–377 (1) and AceSat02–750 (2) repeats. M: marker.
PCR with AceSat01–377 and AceSat02–750 primers with genomic DNA of *A. cepa* resulted in ladder-like PCR products (Fig. 2B, S2 and S3), confirming the tandem organization of these repeats in *A. cepa*. The AceSat02–750 clone was used as a probe for southern hybridization of genomic DNA. After *EcoRI* and *XbaI* digestion, AceSat02–750 revealed a ladder-like pattern (Fig. 3 and S3B), confirming its organization in the form of tandemly repetitive sequences.

**Chromosome localization of AceSat01–377 and AceSat02–750.** SatDNA sequences were located at heterochromatic regions, which appear not only at the centromeric and subtelomeric regions of the chromosomes, but also at intercalary positions. FISH with AceSat01–377, AceSat02–750 and 45S rDNA was carried out on the metaphase chromosomes of *A. cepa* to investigate their chromosomal distribution (Figs 4 and 5).
45S rDNA localized on three chromosomes (Fig. 4A and S5). In most cases in A. cepa, there are two pairs 45S rDNA loci reported. However, in A. cepa, two, three, or four loci of 45S rDNA would be expected, which might be due to mobility of NOR. The AceSat01–377 clone hybridized at sub-terminal regions at both ends of only one pair of chromosomes, and labeled only one end of rest of chromosomes except two chromosomes (Figs 4B and 5). The heterozygous signals of AceSat01–377 for each chromosome are strong enough and it is unlikely one chromosome is unlabeled while the homologous is labeled due to technique issue. In addition, the heterozygous of rDNA and repeats are reported in Allium cepa, Vicia faba and many other plants. AceSat02–750 hybridized to sub-terminal regions on three pairs of chromosomes. AceSat01–377 and AceSat02–750 are co-localized on three chromosome arms (Figs 4B and 5). Peška et al. (2019) also analyzed distribution pattern of similar repeats on chromosomes by FISH. The AcepSAT356 and AcepSAT750 are similar to present AceSat01–377 and AceSat02–750, respectively, but the differences of FISH pattern in these two studies suggested the cultivar difference of these two repeats in A. cepa. AceSat02–750 occurred distal to AceSat01–377 at one end of three chromosomes (Figs 4B and 5). Taking together, we could identify 6 out of 16 chromosomes in A. cepa combining these three probes (Fig. 5). However, we failed to get FISH signals with other candidates (Table S1). Possibly, they are clustered in small groups, which are not sufficient to yield unambiguous FISH signals.

Data availability
All the data pertaining to the present study have been included in tables and/or figures in the present manuscript and the raw reads of sequencing data have been uploaded on the NCBI SRA database. The output of RepeatExplorer archive has been uploaded on the Figshare.com (https://figshare.com/s/b5ad97d66269b0369bc).

Received: 27 February 2019; Accepted: 26 October 2019;
Published online: 14 November 2019

References

1. Leitch, A. R. & Leitch, I. J. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* **194**, 629–648 (2012).
2. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science.* **326**, 1112–1115 (2009).
3. Fritsch, R. M. & Friesen, N. Evolution, domestication, and taxonomy. In: Rabinowitch, H. D. and Currah, L., Eds, *Allium Crop Science:* Recent Advances, CAB Publishing, Wallingford, U. K., pp5–30 (2002).
4. Ricoch, A., Yockteng, R., Brown, S. C. & Nadot, S. Evolution of genome size across some cultivated *Allium* species. *Genome.* **48**, 511–520 (2005).
5. Flavell, R. B., Bennett, M. D., Smith, J. B. & Smith, D. B. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem Genet.* **12**, 257–269 (1974).
6. Kim, S., Park, J. Y. & Yang, T. J. Characterization of three active transposable elements recently inserted in three independent *DFR-A* alleles and one high-copy DNA transposon isolated from the pink allele of the *ANS* gene in onion (Allium cepa L.). *Euphytica.* **120**, 1027–1037 (2015).
7. Barnes, S. R., James, A. M. & Jamieson, G. The organisation, nucleotide sequence, and chromosomal distribution of a satellite DNA from *Allium cepa.* *Chromosoma.* **92**, 185–192 (1985).
8. Do, G. S., Seo, B. B., Yamamoto, M., Suzuki, G. & Mukai, Y. Identification and chromosomal location of tandemly repeated DNA sequences in *Allium cepa.* *Genes Genet. Syst.* **76**, 53–60 (2001).
9. Mancia, F. H. et al. Distribution of various types of repetitive DNAs in *Allium cepa* L. based on dual color FISH. *Hortic. Environ. Biotechnol.* **56**, 793–799 (2015).
10. Fajkus, P. et al. *Allium* telomeres unmasked: the unusual telomeric sequence (CTCCGTTATGGG)n is synthesized by telomerase. *Plant J.* **85**, 337–347 (2016).
11. Peška, V., Mandáková, T., Ihradská, V. & Fajkus, J. Comparative dissection of three giant genomes: *Allium cepa, Allium sativum,* and *Allium ursinum.* *Int. J. Mol. Sci.* **20**, E735 (2019).
12. Kelly, L. J. & Leitch, I. J. Exploring giant plant genomes with next-generation sequencing technology. *Chromosoma Res.* **19**, 939–953 (2011).
13. Novák, P., Neumann, P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *Bioinformatics.* **11**, 378 (2010).
14. Novák, P., Neumann, P., Pech, J., Steinhauser, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of euukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics.* **29**, 792–793 (2013).
15. Mehrotra, S. & Goyal, V. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genom Proteom Bioinf.* **12**, 164–171 (2014).
16. Wicker, T. et al. A unified classification system for euukaryotic transposable elements. *Nat Rev Genet.* **8**, 973–982 (2007).
32. Pich, U., Fritsch, R. & Schubert, I. Closely related
31. Kirov, I. V ., Kiseleva, A. V ., Laere, K. V ., Roy, N. V . & Khrustaleva, L. I. Tandem repeats of
33. Schubert, I. & Wobus, U.
29. Ruiz-Ruano, F . J., López-León, M. D., Cabrero, J. & Camacho, J. P . M. High-throughput analysis of the satellitome illuminates satellite
28. Heckmann, S.
27. Kiseleva, A. V ., Kirov, I. V . & Khrustaleva, L. I. Chromosomal organization of centromeric Ty3/gypsy retrotransposons in
26. Hertweck, K. L. Assembly and comparative analysis of transposable elements from low coverage genomic sequence data in
25. Bennetzen, J. L. & Wang, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes.
24. Kelly, L. J.
23. Sun, C. et al. LTR retrotransposons contribute to genomic gigantism in plesodontid salamanders. Genome Biol. Evol. 4, 168–183
22. Metcalfe, C. J., Filee, J., Germon, I., Joss, J. & Casane, D. Evolution of the Australian lungfish (Neoceratodus forsteri) genome: A major role for CR1 and L2 line Elements. Mol. Biol. Evol. 29, 3529–3539 (2012).
21. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 38, e131 (2010).
20. Weiss-Schneeweiss, H., Leitch, A. R., McCann, J., Jang, T. S. & Macas, J. Employing next generation sequencing to explore the repeat landscape of the plant genome. In: Hörandl, E. & Appelhans, M. S. (eds) Next-generation sequencing in plant systematics. International Association for Plant Taxonomy Publishing, Germany, pp 1–25 (2015).
19. Li, L., Arumuganathan, K., Gill, K. S. & Song, Y . Flow sorting and microcloning of maize chromosome 1. Flow sorting and microcloning of maize chromosome 1. Hereditas. 141, 55–60 (2004).
18. He, S. B. et al. Comparative analysis of genome-wide chromosomal histone modification patterns in maize cultivars and their wild relatives. PLoS One. 9, e97364 (2014).
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
16. Marques, M. P., Sutherland, G. L. & Munn, J. M. The impact of the DNA barcoding approach on the assessment of species diversity and population size of a ground squirrel. J. Hered. 101, 222–230 (2010).
15. Vila, J., Guzmán-Díaz, L., Pérez-Rams, E., Galindo-Rivera, L. & Azcón-Bieto, J. A. The Neogene diversity and biogeography of the Mexican Vachellia floridana complex (Leguminosae, Faboideae). J. Biogeogr. 38, 1918–1929 (2011).
14. Hedges, S. B., Kumar, S. & Ryan, L. A. Dates and rates: lessons from phylogenetic tree dating. Trends Ecol. Evol. 15, 333–339 (2000).
13. Barbrook, P. M., Thomas, C. D. & Reddell, H. J. A. Comparative genomics of plant genomes and the emerging field of genomics-assisted agriculture. Curr. Opin. Biotechnol. 26, 197–203 (2014).
12. Bennett, M. B. & Tatham, A. L. The use of genomics in plant breeding: A review. Euphytica. 171, 1–7 (2010).
11. Belshaw, R. & Siddall, M. E. The DELTRAN toolbox: from DNA sequences to phylogenetic trees. Mol. Phylogenet. Evol. 47, 573–580 (2008).
10. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 1–28 (2005).
9. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 349–373 (2005).
8. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 1–28 (2005).
7. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 349–373 (2005).
6. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 1–28 (2005).
5. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 349–373 (2005).
4. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 1–28 (2005).
3. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 349–373 (2005).
2. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 1–28 (2005).
1. Nason, D. C., Shen, H. S. & Brown, W. The phylogenetic relationships of the dandelion tribe (Compositae: Asteroideae). J. Syst. Evol. 43, 349–373 (2005).

Acknowledgements
This work was supported by National Natural Science Foundation of China (NSFC Nos. 31571265 and 31801026). Computational resources were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure.

Author contributions
J.F., Z.Q., L.L. conceived the project and designed this research; J.F., F.G., H.Z. performed the experiments with assistance from L.M., J.W., M.Y., L.L. and X.Z. All authors analyzed the data and discussed the results; J.F., H.Z., L.M., L.L. and Z.Q. wrote the article. All the authors have approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-52995-9.

Correspondence and requests for materials should be addressed to Z.Q. or L.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019