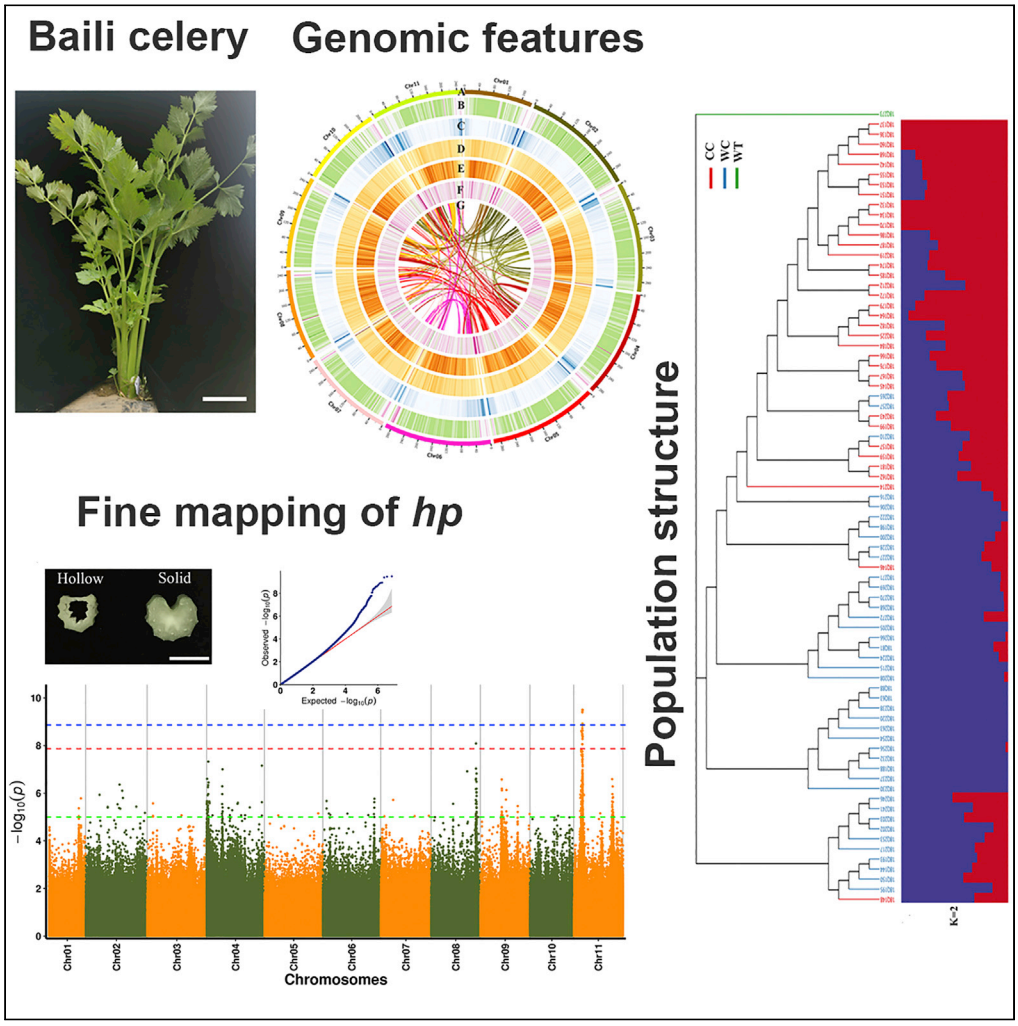
iScience

CellPress
OPEN ACCESS

Article

Loci underlying leaf agronomic traits identified by re-sequencing celery accessions based on an assembled genome



Qing Cheng, Liang Sun, Han Qiao, ..., Haoran Wang, Wencai Yang, Huolin Shen

shl1606@cau.edu.cn

Highlights

Provides a high-quality assembly of Baili celery genome

Re-sequencing and population structure analysis of 79 celery accessions

Performed genome-wide association studies of leaf agronomic traits of celery

Fine mapping the hollow petiole (*hp*) loci in an 807.6-kb region on chromosome 11

Cheng et al., iScience *25*, 104565
July 15, 2022
https://doi.org/10.1016/j.isci.2022.104565

iScience

CellPress
OPEN ACCESS

Article

Loci underlying leaf agronomic traits identified by re-sequencing celery accessions based on an assembled genome

Qing Cheng,[1] Liang Sun,[1] Han Qiao,[1] Zixiong Li,[1] Mingxuan Li,[1] Xiangyun Cui,[1] Wenjie Li,[1] Sujun Liu,[1] Haoran Wang,[1] Wencai Yang,[1] and Huolin Shen[1,2,*]

SUMMARY

Celery is one of the most popular vegetables in the world. The main edible parts of celery are the leaf blade and petiole. The celery petiole is usually green, red, or white, with a hollow or solid pith. However, the loci/genes controlling these petiole-related traits have not been reported. In this study, we present a chromosome-level celery genome assembly with a total size of 3.339 Gb. Simultaneous bursts of long-terminal repeats (78.43%) contributed greatly to the large genome size. Re-sequencing and population structure analysis of 79 celery accessions revealed that they could be divided into Chinese celery and Western celery. By combining genome-wide association studies (GWAS) and mapping data, we located the hollow petiole (*hp*) loci in an 807.6-kb region on chromosome 11. This study provides valuable resources for genetic research on celery and is also helpful for the identification and cloning of genes controlling leaf agronomic traits in celery.

INTRODUCTION

Celery (*Apium graveolens* L., 2n = 2x = 22) is an annual or biennial herbage species that belongs to the Apiaceae family and originates from the Mediterranean and Middle East. Celery was initially domesticated as a medicinal plant, and later became a popular vegetable worldwide (Sturtevant, 1886; Li et al., 2018). It was introduced to China during the Han Dynasty (second century B.C.), and the current area under celery cultivation is ~650,000 hectares, which is almost 3% of the total Chinese vegetable planting area and produces approximately 20 million tons of celery each year. Celery is rich in nutrients, such as vitamins (especially vitamins K and C), apigenin, carotenoids, and cellulose (Fazal and Singla, 2012; Li et al., 2014; Dianat et al., 2015), and is also a good source of flavonoids, volatile oils, and antioxidants (Sowbhagya et al., 2010; Sowbhagya, 2014). In addition, celery can be utilized in the chemical and medical industries (Nagella et al., 2012; Kooti et al., 2014).

Based on their edible parts, celery can be classified as celeriac (*Apium. graveolens* var. *rapaceum*) and leaf celery (*Apium graveolens* var. *dulce*). Celeriac is also called root celery or turnip-rooted celery, and its expanded spherical root is the major edible part. Leaf celery is often abbreviated as celery, of which the edible parts are the leaf blades and petioles. Celeriac is mainly grown in Europe and the USA but is barely cultivated in China. Leaf celery is grown in Western Europe, the USA, Japan, and China (Rožek, 2013; Salehi et al., 2019). In China, based on its origin and morphological features, leaf celery is generally classified as Chinese celery (also known as local celery) or Western celery. Chinese celery has been grown and selected in China for a long time and has the following features: a long, slender, and hollow petiole (*hp*) with a white or light green color, easy to cook, and strong fragrance. Western celery, introduced in China in modern times, is generally tall and dark green, difficult to bolt, develops short, thick, solid petioles, and emits light flavors.

In most plants, the petiole connects the leaf and the stem. In some leafy vegetables, the petiole is not only the major edible part but also plays a stem-like function. In addition, the petiole is also an important channel for transporting water, nutrients, and photosynthetic products and plays an important role in maintaining leaf angle and plant architecture (Tsukaya et al., 2002; Kozuka et al., 2005, 2010). In celery, petioles can

[1]Beijing Key Laboratory of Growth and Developmental Regulation for Protected Vegetable Crops, Department of Vegetable Science, College of Horticulture, China Agricultural University, Beijing 100193, China

[2]Lead contact

*Correspondence: shl1606@cau.edu.cn

https://doi.org/10.1016/j.isci.2022.104565

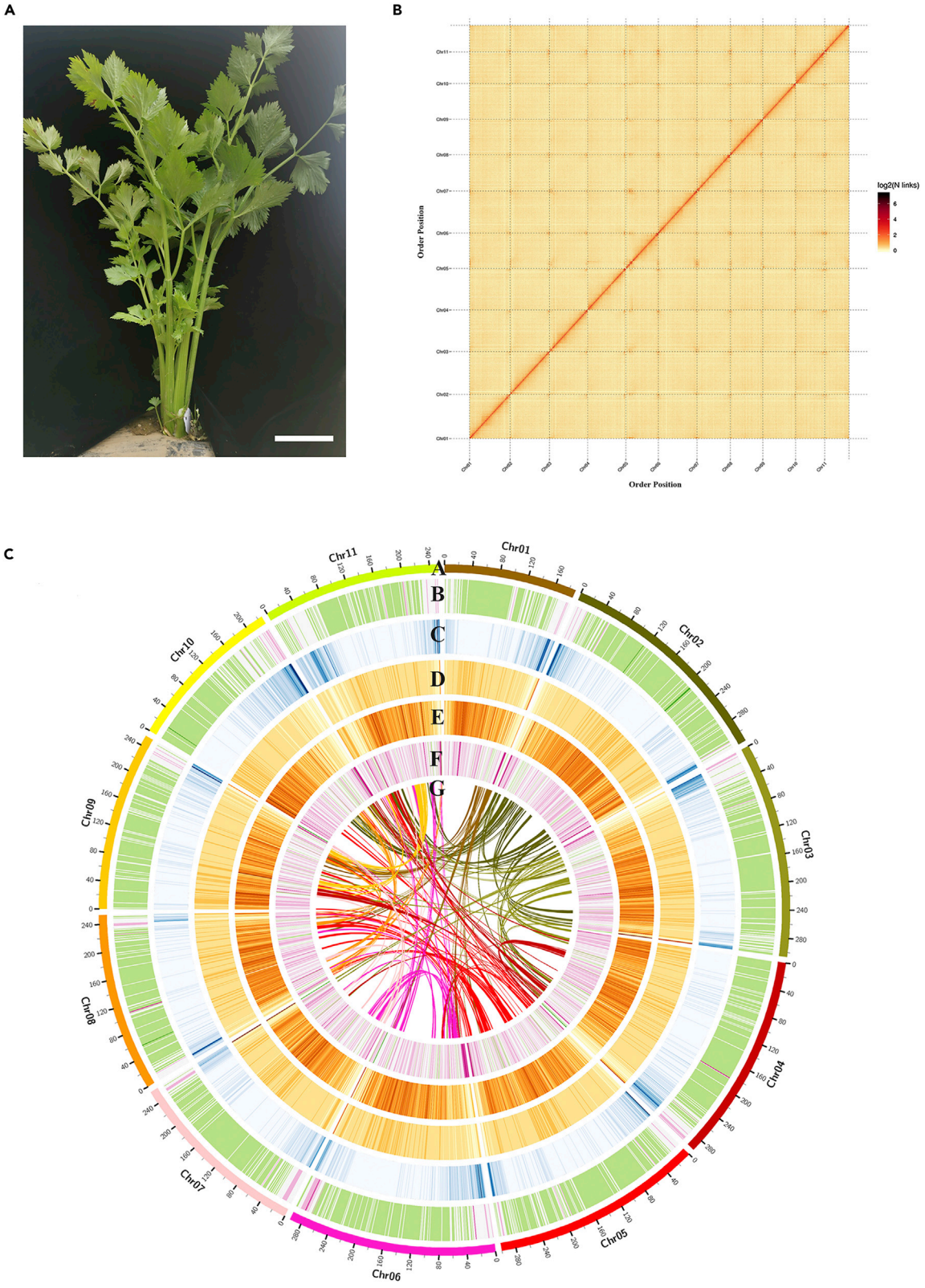

A
B
Order Position
Order Position
log2(N links)
C
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
A
B
C
D
E
F
G

Figure 1. Morphology, Hi-C map, and chromosomal features of celery genomes

(A) The morphology of Baili.

(B) Hi-C map showing genome-wide all-by-all interactions between chromosomes.

(C) Landscape of genome assembly and annotation of Baili. Tracks from outside to the inner correspond to 11 chromosomes of celery; GC content density; gene density; DNA repeats density; density distribution of *Gypsy*; density distribution of *Copia*; relationship between syntenic blocks, as indicated by lines. Each line represents a syntenic block; block size = 3 kb. See also Tables S1–S4.

have a hollow or solid pith. Petiole structure is an important horticultural trait that influences customer decisions. However, the loci/genes that control this trait remain unclear.

Among the Apiaceae species, the carrot and coriander genome has been reported with N50 scaffold lengths of 12.7 Mb and 2.15 Mb, respectively (Iorizzo et al., 2016; Song et al., 2020). Two versions of the celery genome have been reported in recent years, and the sequencing variety was Q2-JN11 and Ventura, respectively (Li et al., 2020; Song et al., 2021). Q2-JN11 belongs under Chinese celery derived from "Jinnan Shiqin," while the Ventura belongs under Western celery. With improvements in sequencing and assembly technology, the genome size and N50 scaffold length of Ventura were found to be much higher than those of Q2-JN11. All previously published genomes of the Apiaceae species provide valuable information for researchers.

In this study, we generated a high-quality chromosome-scale genome assembly of the celery inbred line, Baili. Based on this genome, re-sequencing of 79 celery accessions, including Chinese and Western celeries, was performed; and later genome-wide association studies (GWAS) for several leaf agronomic traits were conducted. The data obtained in this study not only provide a better understanding of the genetic evolution and divergence of the Chinese and Western celeries but also clarify the loci/genes underlying the *hp*. The celery genome assembly in this study will provide valuable resources for facilitating celery genetic research and improvement as well as for studying the evolution and speciation of Apiaceae species. Moreover, the identification of *hp* loci is not only useful for celery genetic improvement but also deepens our understanding of the molecular mechanisms underlying the *hp*.

RESULT

Genome sequencing, assembly, and annotation

Baili (*Apium graveolens* L.), a highly inbred celery line that belongs to the group of Western celery and derived from California celery via single plant selection, was used for genome sequencing. The characteristics of Baili were similar to those of most Western celery: tall and light green, with developed thick and solid petioles (Figure 1A). The genome was sequenced using Illumina HiSeq, PacBio Sequel, and chromosome conformation capture interaction mapping (Hi-C) platforms. A total of 173.81 Gb (53.24×) Illumina paired-end short reads (270 bp) were obtained and used to estimate the heterozygosity ratio of the sampled individuals (0.02%). Based on the 21-mer depth distribution of the Illumina reads (Table S1; Figure S1), the celery genome size was estimated at 3.26 Gb, which was close to the estimates by flow cytometry (3.10 Gb) (Figure S2). A total of 181.61 Gb (54.40×) of clean subreads with a mean read length of 14.34 kb were obtained using the PacBio Sequel system and then used to assemble a 3.338 Gb genome with a contig N50 of approximately 1.03 Mb (Tables S2 and S3). With the aid of Hi-C interaction data, the genome was assembled to 3.339 Gb; 96.59% (3.22 Gb) of the assembled genome was anchored onto 11 pseudochromosomes, and the scaffold N50 was approximately 258.97 Mb (Tables 1 and S4). The genome size, contig, and scaffold N50 values of the Baili assembly genome and Ventura assembly genome were much higher than the Q2-JN11 assembly genome (Li et al., 2020; Song et al., 2021, Table S5); and the genome size, contig, and scaffolds N50 values, as well as the comparison of genomic structure, showed that our assembly genome had a similar quality to the Ventura assembly genome (Table S5), indicating that both the Baili and Ventura assembly genomes were of high quality.

To evaluate the quality of the celery genome assembly, Illumina and RNA-seq reads were first aligned against the genome, and the proper mapping rates were 95.36% and 78.49%, respectively (Tables S6 and S7). Next, the core eukaryotic genes (CEGs) and BUSCO database were searched, and the majority of the CEGs (95.2%) and genes in the BUSCO dataset (97.03%) were identified (Table S8). Furthermore, based on the high long terminal repeat (LTR) assembly index (LAI), our celery genome LAI reached 11.31, indicating high quality of our assembly. Finally, a heatmap was drawn with the Hi-C data, and all bins could be clearly classified into 11 pseudochromosomes (Figure 1B). Taken together, these evaluations

Table 1. Statistics for the celery genome

Assembly Feature	Number	Length (bp)	Percentage (%)
Assembled scaffold sequences (>1 kb)	5,358	3,339,076,396	–
N50 scaffold	–	258,965,703	–
N90 scaffold	–	412,247	–
Max scaffold	–	315,280,121	–
Assembled contig sequences (>1 kb)	11,246	3,338,485,196	99.98
N50 contig	–	600,000	–
N90 contig	–	179,658	–
Max contig	–	3,697,775	–
GC content	–	–	35.86
Chromosome	11	3,224,725,386	96.59
Anchored and oriented scaffolds	5,923	2,948,500,488	91.43

indicate that our genome assembly was of high quality, and the overall completeness was acceptable at the chromosome scale.

A total of 32,599 genes were predicted in our assembly genome, which is similar to the two previous celery assemblies (Figure 1C; Table S5). Among these predicted genes, 32,156 (98.64%) shared homology with the annotated genes (Table S9) and 24,082 (73.87%) could be supported by RNA-seq data (Figure S3; Table S10). The predicted genes showed average exons and average CDS lengths similar to the Ventura and coriander genomes (Song et al., 2020, 2021) (Table S11; Figure S4). Among the noncoding RNAs, 70 miRNAs, 780 rRNA, and 647 tRNA were predicted (Table S12). The predicted motifs and pseudogenes are listed in Table S12.

Genomic variation between the Baili and Ventura genomes

We conducted genomic collinearity analysis between the Baili (this study) and Ventura genomes (Song et al., 2021) (Figure 2A). Large structural variations (SVs) between the Baili and Ventura genomes were found on Chr03, Chr06, and Chr11 (Figure 2C; Table S13). A total of 952 inversions, 2,789 translocations, and 18,069 duplications were identified in the syntenic blocks between the two genomes (Table S13). The distribution lengths of these SVs are shown in Figure 2B.

Repetitive sequence contributed to the large genome size of celery

Repetitive sequences accounted for the majority (2.90 Gb, ~81.49%) of the celery genome and were 1.2 times that of coriander (*Coriandrum sativum*) (70.59%) and 1.8 times that of carrot (*Daucus carota*) (46%) (Figure 1C; Table S14). Most transposable elements (TEs) belong to the long terminal repeat (LTR) category, with a total length of over 2.79 Gb, accounting for 78.43% of the whole-genome size. Most of the LTRs were *Copia* and *Gypsy* elements, accounting for 36.94% and 23.16% of the whole genome, respectively (Table S14). Therefore, the substantial accumulation of TEs, especially LTRs, contributed greatly to the large genome size of celery.

To trace the history of the greatly expanded repetitive sequences in celery, the insertion times of all LTRs were estimated. A peak of increased insertion activity was found at ~0.35 Mya (Figure 3A), suggesting that the expansion of the celery genome occurred quite recently. In addition, compared to the other two Apiaceae species, the accumulation of LTRs was much higher and faster in celery and coriander (Figure 3A).

The insertion time analyses of *Copia* and *Gypsy* showed that both LTR retrotransposons had the highest insertion activity after Apiaceae species diverged, which made them the most abundant retrotransposons in the celery genome (Figures 3B and 3C). Analysis of the phylogenetic topology of *Copia* and *Gypsy* clades showed that there were many species-specific LTRs, especially in the celery and coriander genomes (Figures S5 and S6).

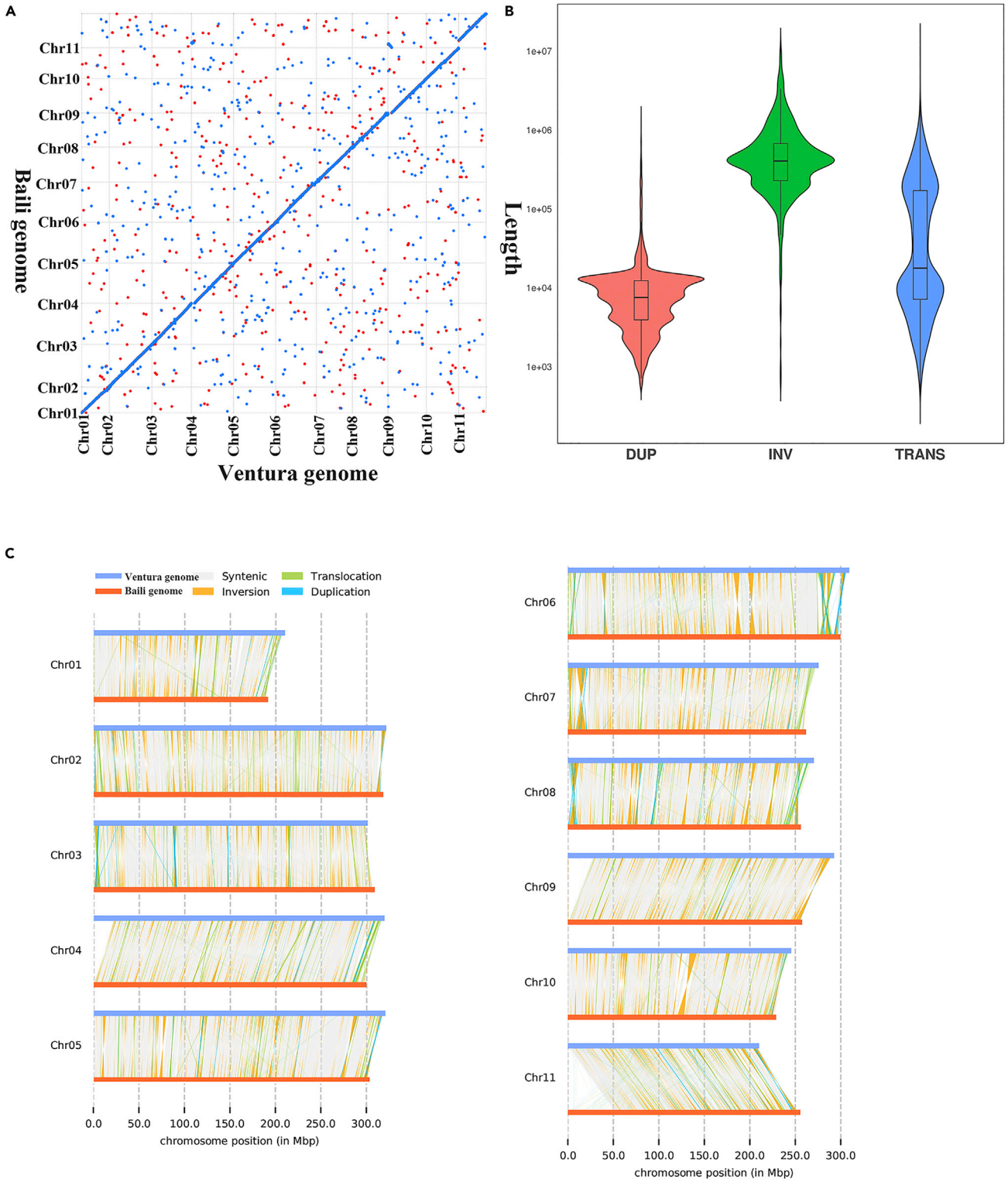


Figure 2. Genomic variation between the Baili and Ventura genomes

(A) Colinearity between the Baili and Ventura genomes. Each dot indicates an aligned region with a length of at least 20 kb.

(B) The distribution length of SVs.

(C) Whole-genome alignment between the Baili and Ventura genomes. See also Tables S5 and S13.

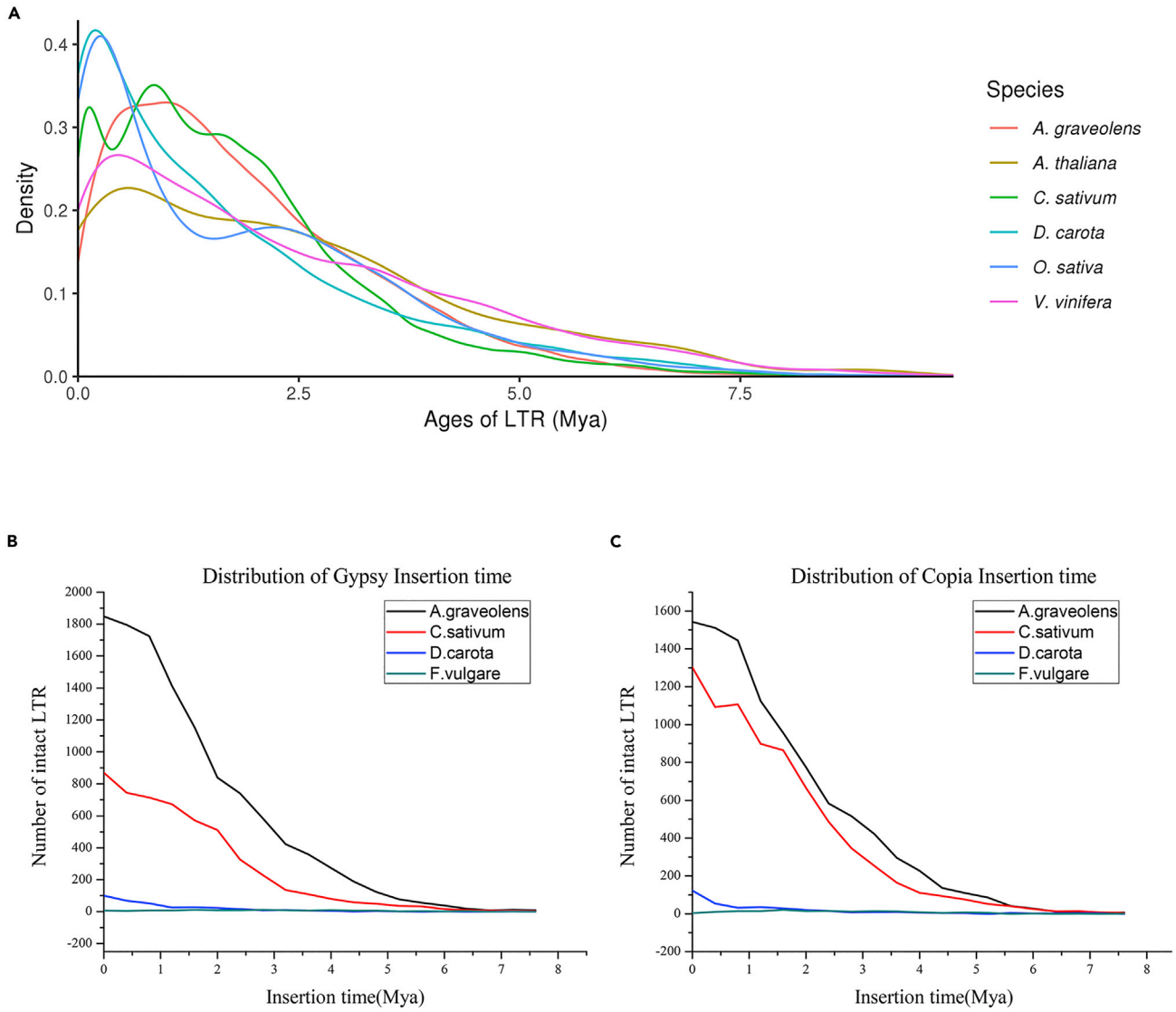


Figure 3. Comparison of transposable elements

(A) distribution of insertion times for LTR retrotransposons in coriander (*C. sativum*), carrot (*D. carota*), fennel (*F. vulgare*), Arabidopsis (*A. thaliana*), rice (*O. sativa*), and celery (*A. graveolens*).

(B) Distribution of insertion times for *Gypsy* in coriander (*C. sativum*), carrot (*D. carota*), fennel (*F. vulgare*) and celery (*A. graveolens*).

(C) Distribution of insertion times for *Copia* in coriander (*C. sativum*), carrot (*D. carota*), fennel (*F. vulgare*), and celery (*A. graveolens*). Mya: million years ago. See also Table S14.

Gene family analysis of celery

To identify the unique and common gene families in celery, celery genes were clustered with the genes of coriander, carrot, sunflower (*Helianthus annuus*), and grape (*Vitis vinifera*). The result showed that there were 14,829 genes in celery that had more than one orthologous gene, and 6,795 gene families were shared between the five species (Figure 4A). Additionally, 1,216 gene families (2,882 genes) were unique to celery. Gene Ontology (GO) analysis showed that 757; 1,418; and 984 of these unique genes were enriched in the cellular component, molecular function, and biological process, respectively; and they were mainly enriched in cell (GO:0044464), catalytic activity (GO:0003824), and metabolic process (GO:0008152) (Figure S7; Table S15). KEGG analysis showed that these unique genes were mainly involved in RNA polymerase, glycosaminoglycan degradation, sphingolipid metabolism, and diterpenoid biosynthesis (Figure S8; Table S16).

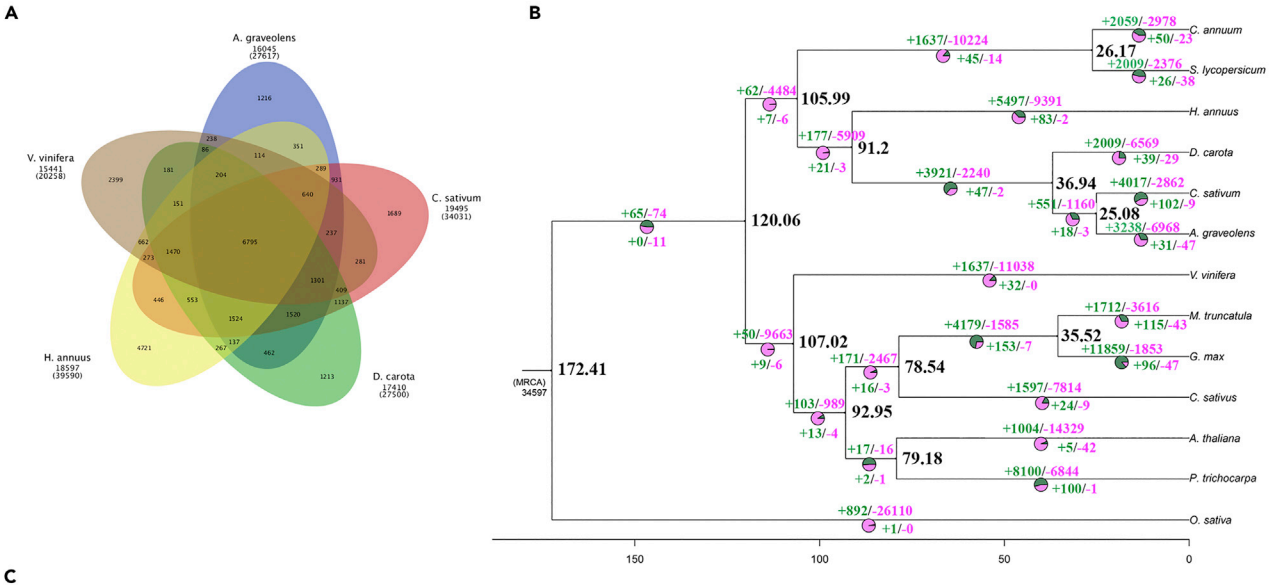

A
A. graveolens
16045
(27617)
V. vinifera
15441
(20258)
C. sativum
19495
(34031)
H. annuus
18597
(39590)
D. carota
17410
(27500)
B
172.41
(MRCA)
34597
120.06
105.99
91.2
107.02
92.95
79.18
78.54
36.94
35.52
26.17
25.08
+65/-74
+0/-11
+62/-4484
+7/-6
+1637/-10224
+45/-14
+2059/-2978
+50/-23
+2009/-2376
+26/-38
+5497/-9391
+83/-2
+177/-5999
+21/-3
+2009/-6569
+39/-29
+3921/-2240
+47/-2
+4017/-2862
+102/-9
+551/-1160
+18/-3
+3238/-6968
+31/-47
+1637/-11038
+32/-0
+50/-9663
+9/-6
+4179/-1585
+153/-7
+1712/-3616
+115/-43
+11859/-1853
+96/-47
+171/-2467
+16/-3
+1597/-7814
+24/-9
+103/-989
+13/-4
+1004/-14329
+5/-42
+17/-16
+2/-1
+8100/-6844
+100/-1
+892/-26110
+1/-0
C. annuum
S. lycopersicum
H. annuus
D. carota
C. sativum
A. graveolens
V. vinifera
M. truncatula
G. max
C. sativus
A. thaliana
P. trichocarpa
O. sativa
150
100
50
0
C
Density
4DTv
0.00
0.25
0.50
0.75
Group
A. graveolens-A. graveolens
A. graveolens-C. sativum
A. graveolens-D. carota
A. graveolens-H. annuus
A. graveolens-V. vinifera
C. sativum-C. sativum
D. carota-D. carota
H. annuus-H. annuus
V. vinifera-V. vinifera
D
Density
Ks
Group
A. graveolens-A. graveolens
A. graveolens-C. sativum
A. graveolens-D. carota
A. graveolens-H. annuus
A. graveolens-V. vinifera
C. sativum-C. sativum
D. carota-D. carota
H. annuus-H. annuus
V. vinifera-V. vinifera

Figure 4. Comparative genomic and phylogenetic relationship analyses

(A) common and lineage-specific gene families in coriander (*C. sativum*), carrot (*D. carota*), Sunflower (*H. annuus*), Arabidopsis (*A. thaliana*), and celery (*A. graveolens*).

(B) Analyses of the divergence time estimation and gene family expansion (green)/contraction (pink).

(C) Distribution of four-fold degenerate sites of the third codons (4DTv) for the percentage of syntenic genes from coriander (*C. sativum*), carrot (*D. carota*), grape (*V. vinifera*), and celery (*A. graveolens*).

(D) Distribution of synonymous substitution rates (Ks) for homologous gene groups. The green and pink numbers above indicate the number of gene families that have expanded and contracted in different species; and the green and pink numbers below indicate the number of gene families that have significantly expanded and contracted in different species ($p < 0.05$).

The annotated genes were clustered with those of carrot, coriander, sunflower, pepper (*Capsicum.annuum*), tomato (*Solanum lycopersicum*), soybean (*Glycine max*), cucumber (*Cucumis sativus*), grape, *Medicago truncatula*, poplar (*Populus trichocarpa*), rice (*Oryza sativa*), and *Arabidopsis thaliana* using OrthoMCL. A total of 16,045 gene families were identified in the celery genome, of which 419 gene families were celery-specific (Table S17; Figure S9). The phylogenetic tree showed that celery diverged from coriander in Apiaceae approximately 25.08 million years ago (Mya) (Figure 4B), and the Apiaceae species appeared nearly 91.2 Mya. Furthermore, there were 3,238 and 6,968 gene families that expanded and contracted in our celery genome, respectively (Figure 4B). Among these expanded and contracted gene families, there were 31 and 47 gene families that expanded and contracted statistically significantly ($p < 0.05$) in the celery genome (Figure 4B).

We also identified chromosome-to-chromosome orthologs between celery and two other species in the Apiaceae family (carrot and coriander). The results showed that in the celery genome, there were 1,450 and 977 syntenic blocks, involving 12,180 and 14,545 genes, compared with carrot and coriander, respectively (Table S18). The degree of collinearity between celery and carrot was 41.19%, whereas that between celery and coriander was 48%, indicating that the three Apiaceae species were very close to each other at the whole-genome level (Figure S10).

Whole-genome duplication in celery

In evolutionary history, most plants have undergone whole-genome duplication (WGD) or polyploidization. Similar to carrot and coriander, there were two peaks in the 4DTv value in the celery genome, indicating that two WGD events occurred in the celery genome lineage (Figure 4C). This result was further confirmed by Ks analysis, in which the Ks values between celery paralogous gene pairs displayed two peaks at 0.618 and 1.1 (Figure 4D).

The Ks of the orthologous genes in celery and coriander peaked at 0.2 and the diverged time between celery and coriander was 25.08 Mya (Figure 4B). We used the substitution rate that was obtained from Song et al. (2021) and calculated that the two WGD events in celery might have occurred at 58-67 and 103-119 Mya, respectively. These results suggest that celery shares WGD with Apiaceae.

Genomic variations and population structure of celery

To explore genetic variations in the celery germplasm, we re-sequenced 79 celery accessions, including 34 Chinese celery accessions, 34 Western celery accessions, 10 hybrid selections from Chinese celery cross Western celery, and 1 wild type (Table S19). A total of 1,376.51 Gb of clean data with an average of ~4× and 97.36% coverage rate of the celery genome was generated with a Q30 of 93.51% (Table S20). After alignment against the celery genome, we identified a total of 17,157,833 high-quality SNPs and 10,662,508 InDels. Among these SNPs, there were 1,660,755 missense variant SNPs and 90,167 stop-gained SNPs (variants causing a STOP codon). In addition, 16,117 and 13,339 SNPs were located at the splice site acceptors and donors, respectively (Table S21). Among these InDels, 3,934,074 InDels caused a frameshift and 36,483 stop-gained InDels caused STOP codons. In addition, 113,700 and 116,014 InDels were located at the splice site acceptors and donors, respectively (Table S22). We also detected SVs among the above 79 celery accessions, and a final set of 496,924 SVs, ranging from 30 bp to 5 Mb, were identified, which included 15,496 insertions, 182,127 deletions, 12,337 inversions, 36,128 duplications, and 250,836 translocations (Tables S23 and S24). These data provide valuable resources for celery biology and genetic breeding.

To further infer the population structure of celery, we performed phylogenetic, model-based ADMIXTURE, and principal component analyses (PCA) for the 79 celery accessions. The results showed that the celery accessions could be classified into two groups: the Chinese celery (CC) group and the Western celery

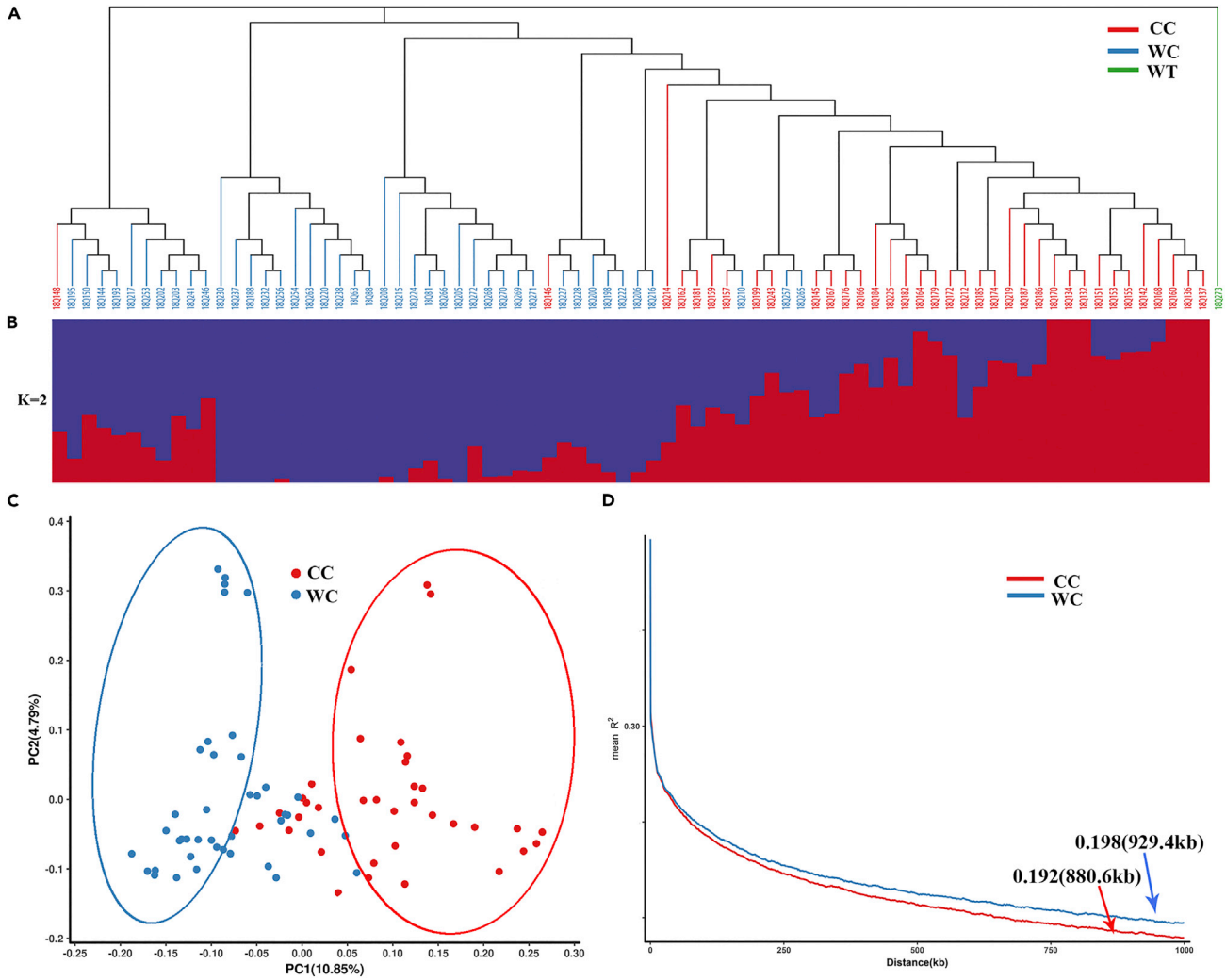


Figure 5. Population structure of the 79 celery accessions

(A) Phylogenetic tree generated from whole-genome SNPs. The 79 accessions could be assigned to the Chinese celery group (CC), Western celery group (WC), and wild type (WT).

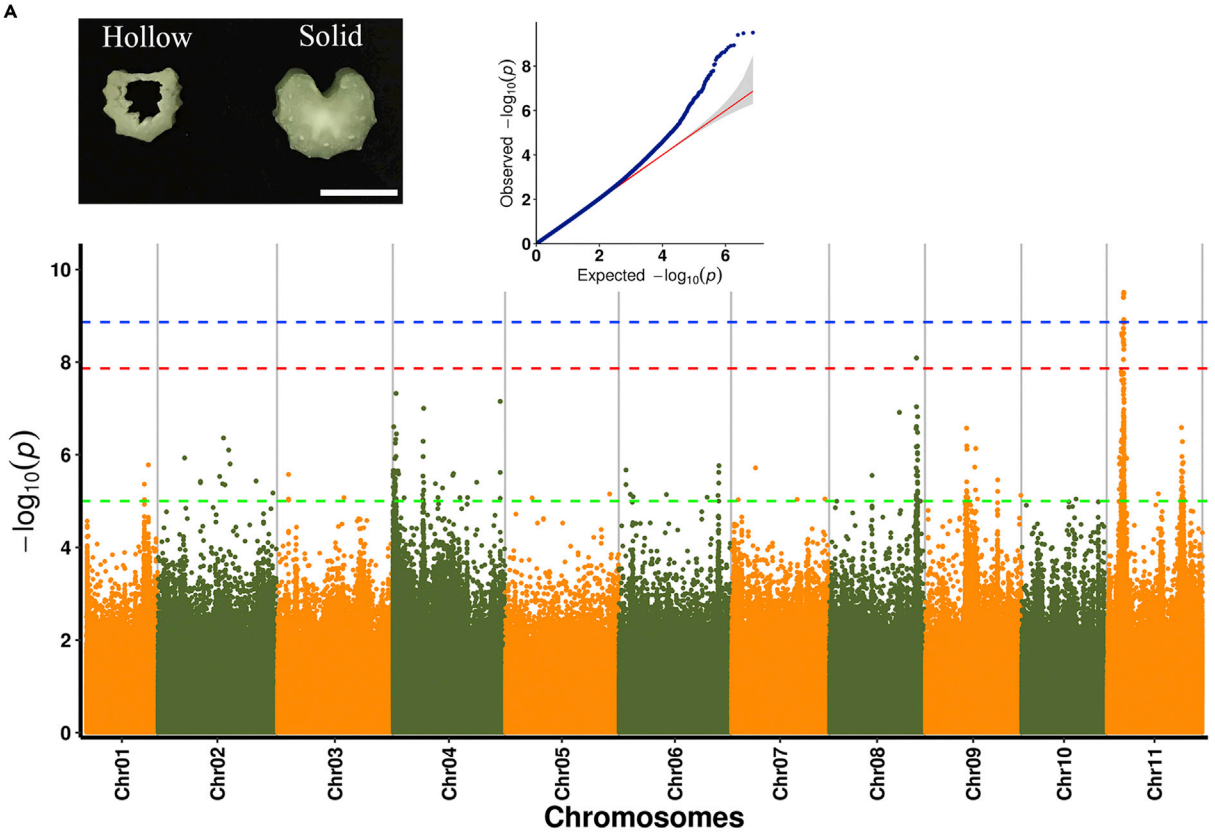
(B) Model-based clustering analysis with the cluster numbers was 2. The y-axis quantifies cluster membership, and the x-axis lists the different accessions. The orders of the 79 accessions on the x axis are consistent with those in the phylogenetic tree.

(C) Principal component analysis of the 79 celery accessions.

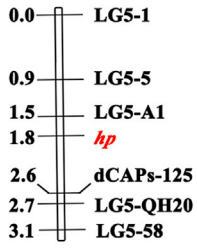
(D) Genome-wide average LD decay estimated from the Chinese celery group (CC) and Western celery group (WC). See also Tables S19, S20, S21, S22, S23, S24, and S25.

group (WC). Forty-one celery accessions were classified into the WC group and thirty-seven celery accessions were classified into the CC group. Wild-type celery formed an outgroup (Figures 5A–5C). Most of the American (13/14) varieties, except for 18Q210; most of the Holland (8/10) varieties, except for 18Q257 and 18Q265; and all of the French varieties (8/8) were classified into the WC group. Structural analysis showed that 18Q210 carried a genetic background closer to Chinese celery than Western celery. The two Holland celery varieties 18Q257 and 18Q265, which were the only two red petiole accessions, were classified into the CC group. All Thailand (2/2) and most of the Chinese (26/33) varieties were classified into the CC group and the 7 remaining Chinese varieties were assigned to the WC group. Among the seven Chinese varieties, 18Q146 was Chinese celery with the unclear origin, and the other six Chinese varieties (18Q148, 18Q214, 18Q224, 18Q232, 18Q246, and 18Q253) were hybrids that came from crosses between the Chinese celery and Western celery (Figures 5A–5C; Table S19). These results indicated that these varieties carried more Western celery genetic backgrounds than Chinese celery. In addition, most Chinese varieties (21/23) carried a Western celery genetic background to a certain degree (Figure 5B).

A



B



C

Phenotype	LG5-QH19	LG5-1	LG5-5	LG5-A1	dCAPs-125	LG5-QH20	LG5-58	LG5-100	Recombinant individuals
C	H	H	H	H	A	A	A	A	46
C	A	A	A	A	H	H	H	H	159
C	H	H	H	H	A	A	A	A	210
C	A	A	A	A	H	H	H	H	233
C	H	H	H	H	A	A	A	A	278

D

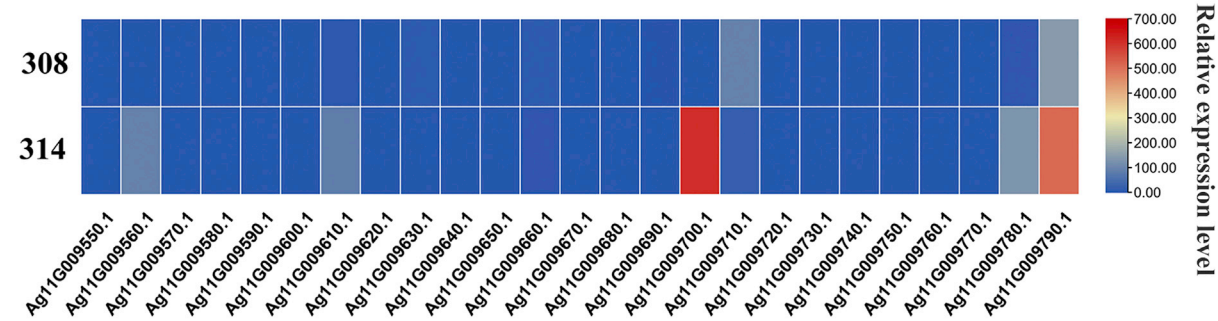


Figure 6. GWAS and mapping of the hollow petiole (*ph*)

(A) Manhattan plots of GWAS for petiole in the 79 celery accessions. Hollow and solid petiole are shown on left, scale bar = 1.0 cm, quantile-quantile plots are shown in right.

(B) Fine mapping of the hollow petiole (*ph*) using F_2 segregating populations.

(C) Identification of genotype of recombinant F_2 individuals. "C" means the hollow petiole phenotype, "A" and "H" means the marker genotype, and "A" means recessive allele, "H" means heterozygous.

(D) RT-PCR analysis of the twenty-five genes in the petiole in lines 308 and 314. See also Tables S28, S29, S30, S33, and S34.

The decay of linkage disequilibrium (LD) with the physical distance between SNPs to half of the maximum values occurred at 880.6 and 929.4 kb in the Chinese celery (r^2 = 0.192) and Western celery (r^2 = 0.198) groups, respectively (Figure 5D).

Divergence between the Chinese celery and Western celery

To obtain genetic insights into Chinese celery and Western celery, we analyzed nucleotide diversity (π). The π values of Chinese celery and Western celery were 6.96×10^{-4} and 5.61×10^{-4}, respectively (Table S25), which is consistent with the result of Watterson estimator analysis (θW (Chinese celery) = 4.79×10^{-4}, θW (Western celery) = 4.55×10^{-4}). Notably, the π values of Western celery were lower than those of Chinese celery, and the LD decay of Western celery was higher than that of Chinese celery, suggesting that Western celery may have experienced a more severe bottleneck during domestication. In addition, both Chinese and Western celery had quite low π values, suggesting that celery had low genetic diversity and a narrow genetic background, and the genetic relationship between the Chinese and Western celery was close.

It is well known that both Chinese and Western celery are cultivated species. Geographically, Western celery is cultivated worldwide, whereas Chinese celery is concentrated in China. The major morphological difference between Chinese and Western celery is the plant architecture. Generally, Chinese celery plants are small and light (0.25-0.5 kg each), and their petioles are slender, leading to loose plant architecture. In contrast, plants of Western celery are large and heavy (1.5-2.5 kg each), and their petioles are thick, leading to compact plant architecture (Figures S11A and S11B). To identify the loci underlying these plant architecture-related traits, we measured the population fixation statistics (F_{ST}) of SNPs in Chinese and Western celery. The average F_{ST} value between Chinese and Western celery was estimated to be 0.122, and the top 5% had $F_{ST} \geq 0.315$ (Table S26). These results indicate moderate population differentiation between the two subspecies (Wright, 1978). Based on F_{ST}, 366 divergent genomic regions were identified, which included 1,770 predicted genes and covered 5.03% (168.1 Mb) of the genome (Figure S11C; Table S27).

To better understand the function of the genes in divergent genomic regions, we performed GO analysis for these genes. Interestingly, several GO terms related to cell development and cell wall construction-related GO terms were enriched. For example, in the "Biological Process" term, cell growth (GO:0016049), developmental cell growth (GO:0048588), plant-type secondary cell wall biogenesis (GO:0009834), cell differentiation (GO:0030154), and plant-type cell wall organization (GO:0009664) were enriched; in the "Cellular Component" term, plant-type cell wall (GO:0009505), cell wall (GO:0005618) and plant-type vacuole membrane (GO:0009705) were enriched; and in the "Molecular Function" term, cellulose synthase (UDP-forming) activity (GO:0016760) and pectatelyase activity (GO:0030570) were enriched (Table S27).

Identification of genes or loci underlyng the hollow petiole locus in celery

Petiole structure is an important quality trait for both Chinese and Western celery, which affects yield and mouth feel. Celery petioles are either hollow or solid (Figure 6A). In this study, experiments were conducted to identify genes or loci underlying the *hp* locus. From the GWAS results obtained from the 79 celery accessions (Table S19), strong association signals were identified on chromosome 11 (Figure 6A; Table S28). To further validate the above signal, an F_2 segregating population was generated from a cross between *hp* line 308 and solid petiole line 314. A 3:1 segregation ratio for the hollow and solid petioles was discovered in the population (Table S29), indicating that the *hp* is controlled by a single gene and the *hp* is dominant over the solid petiole. Subsequently, SSR markers were developed and screened with a 5 Mb interval. Using these markers, *hp* was narrowed to an 807.6-kb region on chromosome 11 (Figures 6B and 6C). Twenty-five genes were identified in this region (Table S30), and we conducted the expression analysis for these genes, the result showed that *Ag11G009560*, *Ag11G009610*, *Ag11G009700*, *Ag11G009780*, and *Ag11G009790* were highly expressed in the solid petiole line 314, while only *Ag11G009710* was highly expressed in the *hp* line 308. The other genes were barely detected in the petiole 308 and 314 (Figure 6D). Gene annotation showed that *Ag11G009780* encodes a homolog of Arabidopsis FUCOSYLTRANSFERASE 1 (*FUT1*)

(Table S30). FUT1 catalyzes the transfer of fucose from GDP fucose to terminal galactose residues on the side chain of xylan (Rocha et al., 2016). Mutations in *FUT1*, also called the *mur2* mutant, showed a 98% reduction in L-fucose levels (Vanzin et al., 2002), and may have a load-bearing effect on the xyloglucan cellulose network (Ryden et al., 2003). *Ag11G009710* encodes a homolog of Arabidopsis MALATE DEHYDROGENASE (MDH, AT3G47520) (Table S30), and mutations in PLASTIDIAL NAD-DEPENDENT MALATE DEHYDROGENASE can rescue ROS accumulation and PCD phenotypes in *mod1* (Zhao et al., 2018). As we know, the *hp* line 308 undergoes a process from solid to hollow during development. At development, the degradation of cytoplasm was observed, and then the intercellular space between pre-cavity cells appeared; the parenchyma cells that formed the pith collapsed and broke down, resulting in the petiole becoming hollow (Figure S12). It has been reported that the formation of cavities in many plant stems or leaves is caused by PCD, such as the cavities in the stems of sorghum and wheat (Fujimoto et al., 2018; Nilsen et al., 2020), the pith cavity of bamboo (Guo et al., 2019), and the fistular leaves of *Allium fistulosum* (Ni et al., 2015). Therefore, combined with the annotation and expression analysis of these candidate genes, we speculated that *Ag11G009780* and *Ag11G009710* may be strong candidate genes for celery *hp*.

In addition, signals for petiole color (green, white, and red) and leaflet margin (mucronate or obtuse) were also identified to be located on chromosome 4 and chromosome 7 in the GWAS analysis (Figures S13, and S14, S31, and S32).

DISCUSSION

In this study, a high-quality celery genome assembly was obtained, of which the genome size was larger than that of carrot (~6.8 times) and coriander (~1.5 times) (Iorizzo et al., 2016; Song et al., 2020), and was close to that of the newly reported celery Ventura genome (Song et al., 2021).

In our celery genome, repetitive sequences, most of which were LTRs, accounted for the vast majority of the genome (2.90 Gb, ~81.49%) (Table S14), which is similar to the coriander genome (1.50 Gb, 70.59%; Song et al., 2020). In addition, we found that the celery and coriander genomes shared an expansion of the LTR, whereas the carrot and fennel genomes did not (Figure 3), which may be the reason why the genomes of celery and coriander were much larger than those of carrot and fennel.

In addition to LTR expansion, WGD and polyploidization also significantly affect the genome size of angiosperms (Piegu et al., 2006; El Baidouri and Panaud, 2013). In this study, two WGD events were identified in our celery genome (Figures 4C and 4D), which were also observed in coriander and carrot (Iorizzo et al., 2016; Song et al., 2020), and the WGD events may have occurred 58-67 and 103-119 Mya, respectively. However, this differs from the Ventura genome, where WGD events occurred 34-38 and 66-77 Mya, respectively (Song et al., 2021). The earliest WGD (103-119 Mya) was confirmed to be shared by all eudicots and is similar to the γ event (115-130 Mya) in the Ventura genome. The latter WGD (58-67 Mya) is similar to the ω event (66-77 Mya) in the Ventura genome (Song et al., 2021); however, the α event in the Ventura genome was not found in our celery genome, this maybe owing to the expansion of repetitive sequences were occurred at recently (the peak of LTR increased insertion was 0.35 Mya) in Baili celery genome, so the latest WGD (α event) has not yet occurred.

In this study, based on phylogenetic analysis, the 79 investigated celery accessions could be classified into three groups, which was in good agreement with the plant taxonomy. However, two Chinese celery lines, 18Q146 and 18Q148, were assigned to the Western celery group, while three Western celery lines, 18Q210, 18Q257, and 18Q265, were assigned to the Chinese celery group (Figures 5A–5C). As Chinese celery was domesticated from Western celery, it is not surprising that some of the Chinese celery varieties harbored a certain degree of genetic background originating from Western celery. Based on our records, the three Western celery varieties that were assigned to the Chinese celery group originated in the USA and Holland. However, it was unclear whether they were crossed with Chinese celery before or after they were introduced into China. A similar phenomenon was observed by Wang et al. (2011) and Fu et al. (2013, 2014). Traditionally, celery is classified into China and Western celery based on its origin and morphological features. However, based on our results, morphological data and original records were not fully consistent with the genomic analysis, indicating that the classical division method based only on the origin and morphological characteristics is not accurate. However, by analyzing the nucleotide diversity (π) of the 79 celery accessions, we found that the genetic diversity of celery was low, indicating that the genetic background of celery was narrow.

Most researchers accept that Chinese celery is highly likely to be selected and domesticated from Western celery. In this study, phylogenetic analysis supported the above hypothesis. Considering the differences between Chinese and Western celery at the morphological, structural, and physiological levels, it can be presumed that after celery was introduced into China, based on their own preferences, Chinese breeders selected and bred a special type of celery, which has a long, slender, and *hp* with white or light green color, strong fragrance, and is easy to cook. Meanwhile, the long, slender, and *hp* of Chinese celery also affected the plant architecture, which made the petiole of the Chinese celery less erect and stronger than that of the Western celery. In this study, the F_{ST} of SNPs between Chinese and Western celery was calculated using the re-sequenced data of 79 accessions, and a number of divergent genomic regions that may determine the plant architecture were identified. Genes within these regions were enriched in cell development and cell wall construction in GO analysis, suggesting that these genes may be selected during Chinese celery domestication. GWAS was also conducted to map the *hp* trait. An 807.6-kb region on chromosome 11 was identified to harbor *hp* (Figure 6). Twenty-five genes were identified in the *hp* region, among which *Ag11G009780* and *Ag11G009710* may be responsible for the *hp* in celery.

In summary, in this study, we assembled a high-quality celery genome and provided a genomic variation map of celery, which not only deepens our understanding of the genome evolution of Apiaceae species but also provides a genomic framework for germplasm research and celery quality improvement in the future. Moreover, our data are also valuable for the fine-mapping and cloning of genes controlling leaf agronomic traits in celery.

Limitations of the study

We reported a high-quality assembly of celery (Baili) genome, performed GWAS on several important leaf agronomic traits, and identified several genes that are potentially related to *hp*. However, further experiments are needed to determine the final candidate gene that controls the *hp*. Furthermore, the samples used for population structure analysis were inadequate.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Plant materials
- METHOD DETAILS
  - Genome size estimation
  - De novo genome assembly using PacBio reads
  - Chromosome assembly using Hi-C
  - Transcriptome sequencing
  - Gene annotation
  - Comparative genomic analysis
  - Analysis of full-length LTR retrotransposons
  - Sequence alignment and variation calling
  - Phylogenetic and population analysis
  - Selective sweep regions for celery
  - Identification candidate regions related to leaf agronomic traits in celery
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104565.

ACKNOWLEDGMENTS

This research was supported by the Construction of Beijing Science and Technology Innovation and Service Capacity in Top Subjects (CEFF-PXM2019-014207-000032). We also thank professor Tao Lin from China Agricultural University for useful suggestions during the genome research and article editing.

AUTHOR CONTRIBUTIONS

HS and QC conceived and designed the research. HS and QC conducted sample preparation and sequencing. QC, LS, and ZL performed the assembly and annotation. HQ and ZL worked on genome comparative and population genomic analyses. ML, XC, and WL performed fine mapping. SL and HW prepared material for re-sequencing and phenotype investigation. QC and SL wrote the article. WY and HS revised the article. All authors have read and approved the final article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 13, 2021
Revised: April 23, 2022
Accepted: June 6, 2022
Published: July 15, 2022

REFERENCES

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.

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. https://doi.org/10.1016/s0022-2836(05)80360-2.

Attwood, T., and Beck, M. (1994). PRINTS–a protein motif fingerprint database. Protein Eng. *7*, 841–848. https://doi.org/10.1093/protein/7.7.841.

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res. *19*, 2241–2245. https://doi.org/10.1093/nar/19.suppl.2241.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. Genome Res. *14*, 988–995. https://doi.org/10.1101/gr.1865504.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365–370. https://doi.org/10.1093/nar/gkg095.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res. *33*, D212–D215. https://doi.org/10.1093/nar/gki034.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat. Biotechnol. *31*, 1119–1125. https://doi.org/10.1038/nbt.2727.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genom. *7*, 327. https://doi.org/10.1186/1471-2164-7-327.

Chakraborty, M., Baldwinbrown, J.G., Long, A.D., and Emerson, J.J. (2016). Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. *44*, e147. https://doi.org/10.1093/nar/gkw654.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

Cheng, Q., Li, T., Ai, Y.X., Lu, Q.H., Wang, Y.H., Wu, L., Liu, J., Sun, L., and Shen, H. (2020). Phenotypic, genetic, and molecular function of *msc-2*, a genic male sterile mutant in pepper (*Capsicum annuum* L.). Theor. Appl. Genet. *133*, 843–855. https://doi.org/10.1007/s00122-019-03510-1.

Dianat, M., Veisi, A., Ahangarpour, A., and Fathi Moghaddam, H. (2015). The effect of hydro-alcoholic celery (*Apiumgraveolens*) leaf extract on cardiovascular parameters and lipid profile in animal model of hypertension induced by fructose. Avicenna J. Phytomed. *5*, 203–209.

Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M.J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., et al. (2012). The UniProt-GO annotation database in 2011. Nucleic Acids Res. *40*, D565–D570. https://doi.org/10.1093/nar/gkr1048.

Dolezel, J. (2010). Flow cytometric analysis of nuclear DNA content in higher plants. Phytochem. Anal. *2*, 143–154.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

El Baidouri, M., and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol. Evol. *5*, 954–965. https://doi.org/10.1093/gbe/evt025.

Fazal, S.S., and Singla, R.K. (2012). Review on the pharmacognostical & pharmacological characterization of *Apium Graveolens* Linn. Indo Global J. Pharmaceut. Sci. *2*, 36–42.

Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., et al. (2006). Pfam: clans, web tools and services. Nucleic Acids Res. *34*, D247–D251. https://doi.org/10.1093/nar/gkj149.

Fu, N., Wang, P.Y., Liu, X.D., and Shen, H.L. (2014). Use of EST-SSR markers for evaluating genetic diversity and fingerprinting celery (*Apium graveolens* L.) cultivars. Molecules *19*, 1939–1955. https://doi.org/10.3390/molecules19021939.

Fu, N., Wang, Q., and Shen, H.L. (2013). *De novo* assembly, gene annotation and marker development using Illumina paired-end transcriptome sequences in celery (*Apium graveolens* L.). PLoS One *8*, e57686. https://doi.org/10.1371/journal.pone.0057686.

Fujimoto, M., Sazuka, T., Oda, Y., Kawahigashi, H., Wu, J., Takanashi, H., Ohnishi, T., Yoneda, J., Ishimori, M., Kajiya-Kanegae, H., et al. (2018). Transcriptional switch for programmed cell death in pith parenchyma of sorghum stems. Proc. Natl. Acad. Sci. USA *115*, E8783–E8792.

Gough, J., and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res. *30*, 268–272. https://doi.org/10.1093/nar/30.1.268.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. *29*, 644–652. https://doi.org/10.1038/nbt.1883.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. *33*, D121–D124. Database issue. https://doi.org/10.1093/nar/gki081.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321. https://doi.org/10.1093/sysbio/syq010.

Guo, L., Sun, X., Li, Z., Wang, Y., Fei, Z., Jiao, C., Feng, J., Cui, D., Feng, X., Ding, Y., et al. (2019). Morphological dissection and cellular and transcriptome characterizations of bamboo pith

cavity formation reveal a pivotal role of genes related to programmed cell death. Plant Biotechnol. J. *17*, 982–997.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

Haft, D.H., Selengut, J.D., and White, O. (2003). The TIGRFAMs database of protein families. Nucleic Acids Res. *31*, 371–373. https://doi.org/10.1093/nar/gkg128.

Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol. Biol. Evol. *30*, 1987–1997. https://doi.org/10.1093/molbev/mst100.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). PASTEC: an automatic transposable element classification tool. PLoS One *9*, e91929. https://doi.org/10.1371/journal.pone.0091929.

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., Bowman, M., Iovene, M., Sanseverino, W., Cavagnaro, P., et al. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat. Genet. *48*, 657–666. https://doi.org/10.1038/ng.3565.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467. https://doi.org/10.1159/000084979.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457–D462. https://doi.org/10.1093/nar/gkv1070.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354. https://doi.org/10.1038/ng.548.

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O., and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinf. *19*, 189. https://doi.org/10.1186/s12859-018-2203-5.

Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. *44*, e89. https://doi.org/10.1093/nar/gkw092.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. *5*, R7. https://doi.org/10.1186/gb-2004-5-2-r7.

Kooti, W., Ali-Akbari, S., Asadi-Samani, M., Ghadery, H., and Ashtary-Larky, D. (2014). A review on medicinal plant of *Apium graveolens*. AHM *1*, 48–59.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736. https://doi.org/10.1101/gr.215087.116.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinf. *5*, 59. https://doi.org/10.1186/1471-2105-5-59.

Kozuka, T., Horiguchi, G., Kim, G.-T., Ohgishi, M., Sakai, T., and Tsukaya, H. (2005). The different growth responses of the *Arabidopsis thaliana* leaf blade and the petiole during shade avoidance are regulated by photoreceptors and sugar. Plant Cell Physiol. *46*, 213–223. https://doi.org/10.1093/pcp/pci016.

Kozuka, T., Kobayashi, J., Horiguchi, G., Demura, T., Sakakibara, H., Tsukaya, H., and Nagatani, A. (2010). Involvement of auxin and brassinosteroid in the regulation of petiole elongation under the shade. Plant Physiol. *153*, 1608–1618. https://doi.org/10.1104/pp.110.156802.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). Mega X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. *35*, 1547–1549. https://doi.org/10.1093/molbev/msy096.

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B.H., and Orengo, C. (2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. Nucleic Acids Res. *40*, D465–D471. https://doi.org/10.1093/nar/gkr1181.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., et al. (2004). Smart 4.0: towards genomic data integration. Nucleic Acids Res. *32*, D142–D144. https://doi.org/10.1093/nar/gkh088.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189. https://doi.org/10.1101/gr.1224503.

Li, M.Y., Feng, K., Hou, X.L., Jiang, Q., Xu, Z.S., Wang, G.L., Liu, J.X., Wang, F., and Xiong, A.S. (2020). The genome sequence of celery (*Apium graveolens* L.), an important leaf vegetable crop rich in apigenin in the Apiaceae family. Hortic. Res. *7*, 9. https://doi.org/10.1038/s41438-019-0235-2.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, M.Y., Hou, X.L., Wang, F., Tan, G.F., Xu, Z.S., and Xiong, A.S. (2018). Advances in the research of celery, an important Apiaceae vegetable crop. Crit. Rev. Biotechnol. *38*, 172–183. https://doi.org/10.1080/07388551.2017.1312275.

Li, M.Y., Wang, F., Jiang, Q., Ma, J., and Xiong, A.S. (2014). Identification of SSRs and differentially expressed genes in two cultivars of celery (*Apium graveolens* L.) by deep transcriptome sequencing. Hortic. Res. *1*, 10. https://doi.org/10.1038/hortres.2014.10.

Li, R.Q., Fan, W., Tian, G., Zhu, H.M., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., et al. (2010). The sequence and *de novo* assembly of the giant panda genome. Genom. Appl. Biol. *463*, 311–317. https://doi.org/10.1038/nature08696.

Lima, T., Auchincloss, A.H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., et al. (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res. *37*, D471–D478. https://doi.org/10.1093/nar/gkn661.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964. https://doi.org/10.1093/nar/25.5.955.

Marchler-Bauer, A., Lu, S.N., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. *39*, D225–D229. https://doi.org/10.1093/nar/gkq1189.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

Members from National Genomics Data Center (2020). Database resources of the national genomics data center in 2020. Nucleic Acids Res. *48*, D24–D33. https://doi.org/10.1093/nar/gkz913.

Nagella, P., Ahmad, A., Kim, S.J., and Chung, I.M. (2012). Chemical composition, antioxidant activity and larvicidal effects of essential oil from leaves of *Apium graveolens*. Immunopharmacol. Immunotoxicol. *34*, 205–209. https://doi.org/10.3109/08923973.2011.592534.

Ni, X.L., Su, H., Zhou, Y.F., Wang, F.H., and Liu, W.Z. (2015). Leaf-shape remodeling: programmed cell death in fistular leaves of *Allium fistulosum*. Physiol. Plantarum *153*, 419–431.

Nilsen, K.T., Walkowiak, S., Xiang, D., Gao, P., Quilichini, T.D., Willick, I.R., Byrns, B., N'Diaye, A., Ens, J., Wiebe, K., et al. (2020). Copy number variation of *TdDof* controls solid-stemmed architecture in wheat. Proc. Natl. Acad. Sci. USA *117*, 28708–28718.

Ou, S., Chen, J.F., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. *46*, e126. https://doi.org/10.1093/nar/gky730.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067. https://doi.org/10.1093/bioinformatics/btm071.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295. https://doi.org/10.1038/nbt.3122.

Pfeifer, B., Wittelsbürger, U., Ramosonsins, S.E.R., and Lercher, M.J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. *31*, 1929–1936. https://doi.org/10.1093/molbev/msu136.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., and Panaud, O. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. *16*, 1262–1269. https://doi.org/10.1101/gr.5290206.

Prestridge, D.S. (1991). Signal SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. Comput. Appl. Biosci. *7*, 203–206. https://doi.org/10.1093/bioinformatics/7.2.203.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). *De novo* identification of repeat families in large genomes. Bioinformatics *21*, i351–i358.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909. https://doi.org/10.1038/ng1847.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A., Machol, I., Omer, A., Lander, E., and Aiden, E. (2014). A 3D Map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.

Rocha, J., Cicéron, F., de Sanctis, D., Lelimousin, M., Chazalet, V., Lerouxel, O., and Breton, C. (2016). Structure of *Arabidopsis thaliana* FUT1 reveals a variant of the GT-B class fold and provides insight into xyloglucan fucosylation. Plant Cell *28*, 2352–2364. https://doi.org/10.1105/tpc.16.00519.

Rožek, E. (2013). Yielding of leaf celery *Apium graveolens* L. var. secalinum Alef. depending on the number of harvests and irrigation. Mod. Phytomorphol. *3*, 83–86.

Ryden, P., Sugimoto-Shirasu, K., Smith, A.C., Findlay, K., Reiter, W.D., and McCann, M.C. (2003). Tensile properties of Arabidopsis cell walls depend on both a xyloglucan cross-linked microfibrillar network and rhamnogalacturonan II-borate complexes. Plant Physiol. *132*, 1033–1040. https://doi.org/10.1104/pp.103.021873.

Salehi, B., Venditti, A., Frezza, C., Yücetepe, A., Altuntaş, Ü., Uluata, S., Butnariu, M., Sarac, I., Shaheen, S., A. Petropoulos, S., et al. (2019). *Apium* plants: beyond simple food and phytopharmacological applications. Appl. Sci. *9*, 3547. https://doi.org/10.3390/app9173547.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259. https://doi.org/10.1186/s13059-015-0831-x.

She, R., Chu, J.S.C., Wang, K., Pei, J., and Chen, N.S. (2009). GenBlastA: enabling BLAST to identify homologous gene sequences. Genome Res. *19*, 143–149. https://doi.org/10.1101/gr.082081.108.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Song, X., Sun, P., Yuan, J., Gong, K., Li, N., Meng, F., Zhang, Z., Li, X., Hu, J., Wang, J., et al. (2021). The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales. Plant Biotechnol. J. *19*, 731–744. https://doi.org/10.1111/pbi.13499.

Song, X., Wang, J., Li, N., Yu, J., Meng, F., Wei, C., Liu, C., Chen, W., Nie, F., Zhang, Z., et al. (2020). Deciphering the high-quality genome sequence of coriander that causes controversial feelings. Plant Biotechnol. J. *18*, 1444–1456. https://doi.org/10.1111/pbi.13310.

Sowbhagya, H.B. (2014). Chemistry, technology, and nutraceutical functions of celery (*Apium graveolens* L.): an Overview. Crit. Rev. Food Sci. Nutr. *54*, 389–398. https://doi.org/10.1080/10408398.2011.586740.

Sowbhagya, H.B., Srinivas, P., and Krishnamurthy, N. (2010). Effect of enzymes on extraction of volatiles from celery seeds. Food Chem. *120*, 230–234. https://doi.org/10.1016/j.foodchem.2009.10.013.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics *19*, ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080.

Sturtevant, H.L. (1886). History of celery. Am. Nat. *20*, 599–606. https://doi.org/10.1086/274288.

Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. *43*, e78. https://doi.org/10.1093/nar/gkv227.

Tarailo-Graovac, M., and Chen, N.S. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc Bioinformatics. Chapter 4.

Thomas, P.D., Kejariwal, A., Campbell, M.J., Mi, H., Diemer, K., Guo, N., et al. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. *31*, 334–341. https://doi.org/10.1093/nar/gkg115.

Tsukaya, H., Kozuka, T., and Kim, G.T. (2002). Genetic control of petiole length in *Arabidopsis thaliana*. Plant Cell Physiol. *43*, 1221–1228. https://doi.org/10.1093/pcp/pcf147.

Vanzin, G.F., Madson, M., Carpita, N.C., Raikhel, N.V., Keegstra, K., and Reiter, W.D. (2002). The *mur2* mutant of *Arabidopsis thaliana* lacks fucosylated xyloglucan because of a lesion in fucosyltransferase AtFUT1. Proc. Natl. Acad. Sci. USA *99*, 3340–3345. https://doi.org/10.1073/pnas.052450699.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963. https://doi.org/10.1371/journal.pone.0112963.

Wang, S., Yang, W., and Shen, H.L. (2011). Genetic diversity in *Apium graveolens* and related species revealed by SRAP and SSR markers. Sci. Hortic. *129*, 1–8. https://doi.org/10.1016/j.scienta.2011.03.020.

Wang, Y.P., Tang, H.B., Debarry, J.D., Tan, X., Li, J.P., Wang, X.Y., Lee, T., Jin, H., Marler, B., Guo, H., et al. (2012). *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. *40*, e49. https://doi.org/10.1093/nar/gkr1293.

Wright, S. (1978). Evolution and the genetics of populations. In Variability Within and Among Natural Populations (Chicago University Press).

Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.S., Natale, D.A., Vinayaka, C., et al. (2004). PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res. *32*, D112–D114. https://doi.org/10.1093/nar/gkh097.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. *35*, W265–W268. https://doi.org/10.1093/nar/gkm286.

Yang, Z.H. (2007). Paml 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848. https://doi.org/10.1093/bioinformatics/17.9.847.

Zhao, Y., Luo, L., Xu, J., Xin, P., Guo, H., Wu, J., Bai, L., Wang, G., Chu, J., Zuo, J., et al. (2018). Malate transported from chloroplast to mitochondrion triggers production of ROS and PCD in Arabidopsis thaliana. Cell Res. *28*, 448–461.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Re-sequencing data	This study	Genome Sequence Archive : CRA003635
RNA sequence data	This study	Genome Sequence Archive : CRA003635
Genome assembly data	This study	Genome Sequence Archive : GWHBGBL00000000
Software and algorithms		
BUSCO (v2.0)	Simão et al. (2015)	http://busco.ezlab.org
Canu (v1.5)	Koren et al. (2017)	https://github.com/marbl/canu
WTDBG v1.2.8	NA	https://github.com/ruanjue/wtdbg
FaCCon (v0.732)	Chakraborty et al. (2016)	
Pilon (v1.22)	Walker et al. (2014)	http://broadinstitute.org/software/pilon/
CEGMA (v2.5)	Parra et al. (2007)	http://korflab.ucdavis.edu/Datasets
LAI method	Ou et al. (2018).	https://github.com/oushujun/LTR_retriever
LACHESIS software	Burton et al. (2013)	https://github.com/shendurelab/LACHESIS
LTR_FINDER (v1.05)	Xu and Wang (2007)	https://github.com/xzhub/LTR_Finder
RepeatScout (v1.0.5)	Price et al. (2005)	https://github.com/mmcco/RepeatScout
PASTEClassifier	Hoede et al. (2014)	https://urgi.versailles.inrae.fr/download/repet/PASTEC_linux-x64-2.0.tar.gz
RepeatMaskerv4.0.6	Tarailo-Graovac and Chen (2009)	http://www.repeatmasker.org/RMDownload.html
Augustus (v2.4)	Stanke and Waack (2003)	http://bioinf.uni-greifswald.de/augustus/
SNAP (v2006-07-28)	Korf (2004)	http://korflab.ucdavis.edu/Software
GeMoMa (v1.3.1)	Keilwagen et al. (2016), 2018	http://www.jstacs.de/download.php?which=GeMoMa
HISAT2	Kim et al. (2015)	http://daehwankimlab.github.io/hisat2/download/#index
Stringtie (v1.2.3)	Pertea et al. (2015)	https://ccb.jhu.edu/software/stringtie/index.shtml
GeneMarkS-T (v5.1)	Tang et al. (2015)	http://exon.gatech.edu/license_download.cgi
Trinity (v2.1.1)	Grabherr et al. (2011)	https://codeload.github.com/trinityrnaseq/trinityrnaseq/zip/master
PASA (v2.0.2)	Campbell et al. (2006)	https://github.com/PASApipeline/PASApipeline/wiki
EVM (v1.1.1)	Haas et al. (2008)	https://github.com/EVidenceModeler/EVidenceModeler/archive/v1.1.1.tar.gz
tRNAscan-SE (v 1.3.1)	Lowe and Eddy (1997)	https://github.com/UCSC-LoweLab/tRNAscanSE
GenBlastA (v1.0.4)	She et al. (2009)	http://genome.sfu.ca/genblast/download.html
Gene-Wise (v2.4.1)	Birney et al. (2004)	https://www.ebi.ac.uk/~birney/wise2/
OrthoMCL package (v 2.0.9)	Li et al. (2003)	http://orthomcl.org/orthomcl
CAFÉ software (v 4.2)	Han et al. (2013)	http://sourceforge.net/projects/cafehahnlab/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PAML (v 3.15)	Yang (2007)	http://abacus.gene.ucl.ac.uk/software/paml.html
McscanX	Wang et al. (2012)	https://github.com/wyp1125/MCScanX
MUSCLE (v3.8.31)	Edgar (2004)	http://www.drive5.com/muscle/
BWA (v 0.7.17-r1188)	Li and Durbin (2009)	https://github.com/lh3/bwa.git
SAMtools (v1.6-3-g200708f)	Li et al., 2009	http://www.htslib.org/
GATK (v v3.2-2-gec30cee)	McKenna et al. (2010)	https://gatk.broadinstitute.org/hc/en-us
Manta tool	Chen et al., 2016	https://github.com/Illumina/manta/releases/download/v1.6.0/manta-1.6.0.centos6_x86_64.tar.bz2
MEGA X	Kumar et al. (2018)	https://www.megasoftware.net/
admixture	Alexander et al. (2009)	http://software.genetics.ucla.edu/admixture/download.html
EIGENSOFT63 (v.6.0.1)	Price et al. (2006)	https://github.com/DReichLab/EIG
Software plink	Purcell et al. (2007)	https://www.cog-genomics.org/plink/
PopGenome Software	Pfeifer et al. (2014)	https://cran.r-project.org/web/packages/PopGenome/index.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the lead contact, Huolin Shen (shl1606@cau.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The raw sequence data reported in this study have been deposited in the Genome Sequence Archive in National Genomics Data Center, Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, under accession number CRA003635 that are publicly accessible at https://bigd.big.ac.cn/gsa. The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center (Members from National Genomics Data Center, 2020), Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, under accession number GWHBGBL00000000 that is publicly accessible at https://bigd.big.ac.cn/gwh.

All original code has been deposited at Genome Sequence Archive and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Plant materials

The celery inbred cultivar Baili, which was belong to the Western celery and was derived from California celery via single plant selection, and it had self-crossed eight generations when used for genome sequencing. A total of 79 celery samples were re-sequenced in this study, these included 34 Chinese celery accessions, 34 Western celery accessions, 10 hybrid selections from Chinese celery cross Western celery and one wild type (Table S19). All celery accessions were maintained and cultivated grown in the plastic-covered tunnels at China Agricultural University in Beijing, China.

METHOD DETAILS

Genome size estimation

Genomic DNA was extracted from celery using a standard CTAB protocol. We constructed five two paired-end (PE) libraries with insert sizes of 270 bp following the standard Illumina protocols, and then sequenced on the Illumina HiSeq 2500 platform. We used these data to calculate and plot the k-mer frequency distribution. Genome size can be estimated by the formula G = K_num/peak depth (K_num: the total number of k-mers; peak depth: the depth of the major peak) (Li et al., 2010). The K value was 21. After filtering abnormal depth k-mers, we finally obtained a total of 27,144,018,011 k-mers, and the major peak depth was 44.38. The celery genome size was thus estimated to be 3.26 Gb.

About 1 g leaf from 3-months-old celery was used to obtain the cell nucleus, then the cell nucleus DNA was stained with propidium iodide in dark for 20 min (Dolezel, 2010). DNA was quantified by flow cytometry on a FACSCalibur (BD company, America) and the data was acquired and analyzed by CellQuest (BD company, America) and ModFit (Yerity SoftwareHouse), respectively. *Zea mays* was used as an internal reference. The estimated celery genome size was 3.10 Gb.

De novo genome assembly using PacBio reads

For PacBio library construction, celery genomic DNA was sheared to 20 kb using G-tubes, then purified, damage repair, hairpin adaptor ligation, and digestion with exonuclease to remove the damaged DNA and fragments that without adaptors. The target fragments were size-selected by Blue Pippin electrophoresis. Then the PacBio library was sequenced on a PacBio Sequel sequencer (Pacific Biosciences, CA, USA).

For genome assembly, we first used Canu (v1.5) (Koren et al., 2017) to correct the PacBio data, then we used WTDBG v1.2.8 (https://github.com/ruanjue/wtdbg), FaCCon (v0.732) (Chakraborty et al., 2016) and Canu (v1.5) to independently assemble the high-quality PacBio subreads. These packages yielded 2.087 Gb, 3.184 Gb and 3.335 Gb assemblies, with contig N50 values of 79.57 kb, 1.041 Mb and 1.032 Mb, respectively. The assembly by the Canu (v1.5) resulted in the optimal assembly and was used as a reference. Then the most well assembly genome was corrected with the Illumina data using Pilon (v1.22) (Walker et al., 2014) for three times. To evaluate the assembled genome's quality, we first mapped the Illumina data and PacBio data to it using BWA (Li and Durbin, 2009), then mapped core eukaryotic genes (CEGs) using CEGMA (v2.5) with 458 conserved core eukaryotic genes (Parra et al., 2007), and finally applied a BUSCO (v2.0) test (Simão et al., 2015), with 1440 genes in Embryophyta-odb9 to examine its gene content. We conducted the LAI method followed Ou et al. (2018).

Chromosome assembly using Hi-C

We constructed Hi-C fragment libraries from 300 to 700 bp insert size followed Rao et al. (2014) and sequencing through Illumina NovaSeq 6000 with the PE150 model. The adapter sequences of the raw reads were trimmed, and low-quality PE reads were filtered to obtain clean data. Totally we obtained 157.79 Gb (47.27×) clean Hi-C data, which was used for chromosome-level genome assembly. We then evaluated the alignment efficiency and insert length distribution for valid pair fragments and valid interaction pairs using HiC-Pro (Servant et al., 2015). We first correct errors in scaffolds by splitting the scaffolds into 50 kb segments on average. Then we used LACHESIS software (Burton et al., 2013) to assemble the genome, with the parameters set as follow: CLUSTER_MIN_RE_SITES = 103; CLUSTER_MAX_LINK_DENSITY = 1; CLUSTER_NONINFORMATIVE_RATIO = 5; ORDER_MIN_N_RES_IN_TRUN = 60; ORDER_MIN_N_RES_IN_SHREDS = 60. Finally we performed artificial correction of the LACHESIS-assembled results and gap filling or sequence de-duplication to increase the accuracy and completeness of the assembled genome (Servant et al., 2015).

We then visualized the interaction matrix of all chromosomes and heatmaps with a resolution set at 500kb to assess the accuracy of the Hi-C assembly. Additionally, the final genome assembly was also validated using the next-generation sequencing (NGS) short reads, PacBio long reads and RNA sequencing (RNA-seq) reads.

Transcriptome sequencing

RNA was isolated from the different tissues of celery (fresh leaf, petiole, root) following the manufacturer's protocol provided in the TaKaRa MiniBEST Universal RNA Extraction Kit. The integrity, purity, and concentration of the RNA were assessed using an Agilent 2100 Bioanalyzer, a NanoDrop, and a Qubit 2.0. RNA-seq

library was constructed by mixing an equal amount of RNA from the above different tissues following the NEBNext® Ultra™ RNA Library Prep Kit (NEB) following the manufacturer's instructions and then sequenced on an Illumina HiSeq 2500 platform. Eventually, 11.86 Gb of RNA-seq data with Q30 higher than 93.51% were obtained.

Gene annotation

For repeat annotation, we used the structural prediction and the ab initio prediction methods by using LTR_FINDER (v1.05) (Xu and Wang, 2007) and RepeatScout (v1.0.5) (Price et al., 2005) to construct a primary repeat sequence database. And used the PASTEClassifier (Hoede et al., 2014) to classify the primary database, then formed a final repeat sequence database by combing with the Repbase database (Jurka et al., 2005). Finally, we predicted the repetitive sequence by using RepeatMaskerv4.0.6 (Tarailo-Graovac and Chen, 2009) based on the final repeat sequence database.

For protein-coding gene prediction, we first masked and excluded the repeat elements from the genome assembly, then we used ab initio predictions, homology-based gene models and unigene prediction to predict the high quality protein-coding genes. For ab initio prediction, we used the Augustus (v2.4) (Stanke and Waack, 2003) and SNAP (v2006-07-28) (Korf, 2004). Homologous species prediction was using GeMoMa (v1.3.1) (Keilwagen et al., 2016, 2018) based on *Arabidopsisthaliana*, *Apium graveolens* (Ventura) (Song et al., 2021), *Coriandrum sativum*, *Daucus carota* and *Lactuca sativa*. All RNA-Seq reads were initially aligned against the celery genome using HISAT2 (Kim et al., 2015) and assembled into transcripts using Stringtie (v1.2.3) (Pertea et al., 2015), then open reading frames (ORFs) were predicted using GeneMarkS-T (v5.1) (Tang et al., 2015). RNA-seq reads were *de novo* assembled using Trinity (v2.1.1) (Grabherr et al., 2011) and then analyzed using PASA (Campbell et al., 2006). We then used EVM (v1.1.1) (Haas et al., 2008) to integrate these three prediction methods and performed final modifications using PASA (v2.0.2) (Campbell et al., 2006).

For noncoding RNAs, we predicted microRNAs and rRNAs to search the Rfam database (Griffiths-Jones et al., 2005), and used tRNAscan-SE (v 1.3.1) (Lowe and Eddy, 1997) to predicte the tRNA. And we used GenBlastA (v1.0.4) (She et al., 2009) alignment and GeneWise (v2.4.1) (Birney et al., 2004) to predict the pseudogene.

For gene functional annotation, we blasted the predicted genes against the non-redundant protein (NR) (Marchler-Bauer et al., 2011), KOG (Koonin et al., 2004), GO (Dimmer et al., 2012), KEGG (Kanehisa et al., 2016), TrEMBL (Boeckmann et al., 2003), EggNOG (http://eggnog5.embl.de/download/eggnog_5.0/) (Jaime Huerta-Cepas,et al., 2018), SWISS-PROT (http://ftp.ebi.ac.uk/pub/databases/swissprot) (Boeckmann et al., 2003) and Pfam (http://pfam.xfam.org) (Finn RD, 2006) databases using BLAST (v2.2.31) (-evalue 1e-5) (Altschul et al., 1990). For motifs annotation, we used InterProScan (Zdobnov and Apweiler, 2001) by aligning with the PROSITE (Bairoch, 1991), HAMAP (Lima et al., 2009), Pfam (Finn et al., 2006), PRINTS (Attwood and Beck, 1994), ProDom (Bru et al., 2005), SMART (Letunic et al., 2004), TIGRFAMs (Haft et al., 2003), PIRSF (Wu et al., 2004), SUPERFAMILY (Gough and Chothia, 2002), CATHGene3D (Lees et al., 2012), and PANTHER (Thomas et al., 2003) databases.

Comparative genomic analysis

We used the OrthoMCL package (v 2.0.9) (Li et al., 2003) to identify orthologous genes between celery and 11 other plant species, including *Daucus carota*, *Coriandrum sativum*, *Helianthus annuus*, *Capsicum.annuum*, *Solanum lycopersicum*, *Glycine max*, *Cucumis sativus*, *Vitis vinifera*, *Medicago truncatula*, *Populus trichocarpa*, *Oryza sativa*, and *Arabidopsis thaliana*. Gene family expansion and contraction was analyzed using CAFÉ software (v 4.2) (Han et al., 2013) with a probabilistic graphical model. Phylogenetic tree between these 13 plant species was constructed using PHYML (Guindon et al., 2010) based on the 320 single-copy orthologous genes with the parameters (-gapRatio 0.5 -badRatio 0.25 -model HKY85 -bootstrap 1000). Divergence times were estimated using MCMCtree in PAML (v 3.15) (Yang, 2007), based on the predicted divergence time from *A.thaliana* and *P.trichocarpa* (105.97∼107.96 Mya) and *V.vinifera* and *O.sativa* (151.98∼159.97 Mya).

The all-versus-all blastp method (E-value<1e-5) was used to detect paralogous genes in *Apium graveolens*, *Daucus carotaas*, *Coriandrum sativum* as well as orthologous genes in *Apium graveolens-Daucus carota*, *Apium graveolens-Coriandrum sativum* and *Apium graveolens-Vitis vinifera*. Then gene pairs were

detected using McscanX (Wang et al., 2012) and the 4DTv-value of these gene pairs was calculated using the HKY model.

Paralogs within *Apium graveolens* and orthologs between *Apium graveolens* and *Coriandrum sativum* and *Daucus carotaas* were identified using BLASTP (e-value was 1e-10). Then we used MCscanX to analyze chromosome collinearity (Wang et al., 2012) with the following parameters: match_score (k) = 50; genes required to call a collinear block (s) = 5; gap penalty (g) = 1; maximum gaps allowed (m) = 25; and alignment significance = 1e-10.

Analysis of full-length LTR retrotransposons

We used LTR_Finder (v1.0.5) (Xu and Wang, 2007) to *de novo* detect full-length LTR retrotransposons in genome of *Apium graveolens*, *Daucus carota*, *Coriandrum sativum* and *Foeniculum vulgare*. Next, we screened the LTR sequence of scores higher than 6 by using the PS SCAN (Prestridge, 1991), and filtered those overlapped LTR sequence. The LTR retrotransposons protein sequences were aligned using MUSCLE (v3.8.31) (Edgar, 2004), and we built the neighbor-joining (NJ) trees of *Copia* and *Gypsy* superfamilies using MEGA X (Kumar et al., 2018) with the default parameters. The nucleotide distance was estimated using the Kimura two-parameter (K2p) (transition–transversion ratio) criterion by using DistMat software, and the rate of nucleotide substitution was using the dicotyledon mutation rate, which was used 7.3×10^{-9}.

Sequence alignment and variation calling

The 79 celery accessions genomic DNA was extracted from leaves using the CTAB method. Illumina genomic libraries with insert sizes of 300–500 bp were constructed following the manufacturer's instructions, the libraries then sequenced on an Illumina NovaSeq 6000 platform (Illumina Inc., USA) with 150 bp paired-end reads.

To call SNPs, reads of each accession were mapped to the celery reference genome using BWA (v 0.7.17-r1188) (Li and Durbin, 2009) with the default parameters. Then we count the map results by using SAMtools (v 1.6-3-g200708f) (Li et al., 2009). Before calling SNPs and small InDels, we used Picard (http://sourceforge.net/projects/picard/) to filter the MarkDuplicates reads, then used GATK (v v3.2-2-gec30cee) (McKenna et al., 2010) to detect SNPs and small InDels. SNPs were further filtered using the following criteria: (1) we filtered out SNPs that located nearby InDels within 5 bp and adjoined InDels within 10bp; (2) the variant SNPs in 5 bp window should not more than two; (3) an overall quality (QUAL) score of <30; (4) a variant quality by depth (QD) score <2; (5) a mapping quality (MQ) score of <40; (6) a phred-scaled p value (FS) > 60; (7) the other variable filter parameters were used as default parameters. For SV detection, we used Manta tool (Chen et al., 2016) followed the default parameters.

Phylogenetic and population analysis

A subset of 7,629,138 SNPs with a minor allele frequency (MAF) $\geq$ 0.05 and missing rate $\leq$0.2 from 79 celery accessions were used for phylogenetic and population structure analyses. We used MEGA X (Kumar et al., 2018) to build the phylogenetic treewith 1000 bootstrap replicates. And we used admixture (Alexander et al., 2009) to construct the population structure. Cross-validation error was tested for obtaining the most likely K value varying from 1 to 10. In addition, principal component analysis (PCA) was performed with EIGENSOFT63 (v.6.0.1) (Price et al., 2006) using the above SNP data set. Two dimensional coordinates were plotted for the 79 celery accessions. For linkage disequilibrium (LD) analysis,We used the above SNPs to perform LD using Software plink (Purcell et al., 2007), and LD decay was calculated on the basis of the r^2 value and corresponding distance between any two SNPs within a 1000 kb window.

Selective sweep regions for celery

A 100-kb sliding window along with a step size of 10 kb was used to estimate population polymorphisms through the population fixation index (F_{ST}), nucleotide diversity (π) and watterson estimator (θW)between the Chinese celery and Western celery. F_{ST}, π values and θW were calculated at each window using the PopGenome Software (Pfeifer et al., 2014). Sliding windows with the top 5% highest F_{ST} values were selected initially. Then we merged the neighboring windows into one fragment, and if the distance between two fragments was <100 kb, we also merged them into one region. Finally, these merged regions were considered as highly diverged regions between local and Western celery.

Identification candidate regions related to leaf agronomic traits in celery

The agronomic traits in celery were evaluated three times during the winter of 2017, 2018, and 2019 at China Agricultural University in Beijing, China (Table S18). A total of 7,629,138 SNPs with MAF $\geq$0.05 and missing rate $\leq$ 0.2 were used to carry out GWAS, and GWAS were performed using Emmax program (Kang et al., 2010). Finally, the signals with $p < 10^{-8}$ were considered as the significantly and $p < 10^{-9}$ were considered as extremely significant. The methods used for mapping of *ph* loci were followed Cheng et al. (2020), and the primers used in this study were shown in Table S33.

Total RNA was exacted from the young petiole of line 308 and 314 using the Quick RNA isolation Kit (Huayueyang, China), following manufacturer protocol. First-strand cDNA was synthesized using the SuperScript III First-Strand Synthesis System Kit (Huayueyang, China). The primers used for real-time PCR were designed used the Primer 5.0 (http://www.premierbiosoft.com/primerdesign/) and are listed in Table S34. The real-time PCR was performed with a TB Green® Premix Ex Taq™ (Takara, China), following manufacturer's instructions, on an ABI 7500 real-time PCR system. The thermocycling conditions were set as follows: 95°C for 30s, 40 cycles of 95 °C for 5s and 60 °C for 30s, then melt curve. Relative expression values were calculated using the $2^{-\Delta\Delta Ct}$ method.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification and statistical analysis used in the genome sequencing and assembly, genome quality assessment, evolutionary analysis and comparative genome analysis can be found in the relevant sections of the method details.