

A comprehensive analysis of two Chinese cucumber genomes and a mutant population as resources for precision breeding

Jiaxi Han^{1,2,†}, Jingwei Wei^{2,†}, Weiliang Kong³, Weili Miao³, Lidong Zhang¹ , Yuhe Li¹, Jiawang Li¹, Xin Li^{2,*}, Tao Lin^{2,*}  and Hongyu Huang^{1,*}

¹State Key Laboratory of Vegetable Biobreeding, Tianjin Academy of Agricultural Sciences, Tianjin 300192, China

²Beijing Key Laboratory of Growth and Developmental Regulation for Protected Vegetable Crops, College of Horticulture, China Agricultural University, Beijing 100193, China

³Tianjin Kernel Cucumber Research Institute, Tianjin Kernel Agricultural Technology Co., Ltd, Tianjin 300192, China

*Corresponding authors. E-mail: tjsnykxyhgyjs@tj.gov.cn; lintao35@cau.edu.cn; xinl@cau.edu.cn

[†]These authors contributed equally to this work.

Abstract

Cucumis sativus L., commonly known as cucumber, is an important vegetable crop worldwide, with China as the largest producer, particularly of the North and South China types. While extensive genomic research has focused on the North China type, especially the Chinese Long 9930, studies on the South China type remain limited. In this study, we assembled high-quality genomes of two widely cultivated and representative parent varieties: S36 (North China type) and H19 (South China type), and conducted mutagenesis analyses. Comparative genome analysis revealed a large number of structural variants between two North China types and two South China types, with many of the affected genes showing strong homology to known functional loci, potentially contributing to phenotypic divergence. We also constructed an EMS mutant library through the mutagenesis of S36 and identified a gene that encodes chlorophyll oxidase, demonstrating the method's effectiveness for rapid gene discovery. In conclusion, this study provides valuable insights into the classification and evolution of cucumber, highlighting the promising potential of forward genetic approaches in cucumber breeding.

Introduction

Cucumber (*Cucumis sativus* L.) is a key vegetable crop and a valuable model for genetics and genomics research in Cucurbitaceae [1]. Advances in the high-throughput sequencing technologies have accelerated research on cucumbers, particularly on the genetic variants, such as single-nucleotide polymorphisms (SNPs) and small insertions/deletions (InDels), most of which have been identified mainly through the alignment of short sequencing reads to a reference genome [2–4]. However, a single reference genome cannot fully represent the sequence diversity within a species, a key limitation that has hindered the identification of larger structural variants (SVs), which play crucial roles in genome evolution and determine key agronomic traits in crops [5–8]. Therefore, a pan-genome study is essential to comprehensively examine structural variations in the cucumber genome. Insights from a pan-genome study can advance research on cucumber domestication and genes related to vital agronomic traits [5, 9–13].

An ethyl methanesulfonate (EMS) mutant population serves as a valuable tool for identifying new genes, understanding gene functions, and exploring the molecular mechanisms underlying

key agronomic traits [14–16]. Genome assembly, a crucial tool in plant breeding and crop improvement, can be used in combination with an EMS mutant library to discover genes and develop novel breeding materials through forward genetic approaches [17]. For example, a study constructed the EMS mutant population of Chinese cabbage from a double haploid inbred line, A03, and successfully identified two chloroplast-associated genes through forward genetics with the sequenced and assembled A03 genome [15]. Similarly, a high-quality, gap-free telomere-to-telomere genome of the watermelon inbred line G42 was assembled. The EMS mutagenesis protocol was established, which generated 48 monogenic phenotypic mutations. Through these mutations, the study identified two genes responsible for elongated fruit shape and male sterility (CIMS1), both of which are caused by a single base change from G to A [16]. Many genes controlling the key agronomic traits in cucumber, such as fruit length, round leaves, and dwarf plants with short internodes, have been identified and functionally characterized from the EMS mutant populations [18–22]. However, gene functional studies in cucumber using whole-genome resequencing to create genetic variants are lacking.

Cucumis sativus has been categorized into four groups: Indian, Eurasian, East Asian, and Xishuangbanna, and it was introduced

Received: 12 April 2025. Accepted: 16 October 2025. Published: 25 February 2026. Corrected and Typeset: 1 February 2026

© The Author(s) 2025. Published by Oxford University Press on behalf of the Nanjing Agricultural University.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

to China about 2000 years ago [23, 24]. Currently, the North China and South China types are the most widely cultivated cucumber varieties in the country [25]. The North China type is characterized by long fruit, dense warts, and deep and uniformly colored pericarps [26], whereas the South China type typically has short fruits with few warts. However, high-quality genome assemblies of South China type cucumber varieties remain unavailable, and EMS mutagenesis studies on North China type cucumbers have yet to be conducted. To address these gaps, we selected two cucumber varieties with strong commercial performance and comprehensive agronomic traits: S36 (North China type) and H19 (South China type). These varieties are widely used as parental lines in commercial F1 hybrids and have become the core germplasm in cucumber breeding. Considering the importance of an accurate and complete reference genome assembly for genetic and genome-wide studies within species [27], we performed the sequencing of S36 and H19 and assembled two high-quality cucumber genomes, which can serve as a reference for analyzing gene mutations and structural variations. In addition, we conducted EMS mutagenesis to enhance certain traits of S36. We investigated differences in their genomic structure and function by comparing their genomes. By precisely identifying the mutation sites in the EMS mutants and accurately mapping them to the genome, we discovered new genes that regulate cucumber agronomic traits and uncovered the underlying regulatory mechanisms. These findings align with current breeding needs and support the development of an intelligent breeding system based on genomic data.

Overall, this study constructed high-quality genome assemblies of S36 and H19, representing the North China and South China type cucumber varieties, respectively, and conducted a comparative genomic analysis. Type-specific SVs were identified, highlighting genomic differentiation between the two groups. Additionally, EMS mutagenesis was performed on S36 to generate mutants for the identification of genes associated with important agronomic traits, thereby providing genetic resources for the molecular improvement of cucumber.

Results

Genome assemblies of North China type cucumber S36 and South China type cucumber H19

S36, a well-known North China type cucumber, is characterized by its long fruits, dense warts, and lustrous green pericarp. In contrast, H19, a representative South China type, produces short fruits with sparse large warts and uneven pericarp coloration, reflecting pronounced phenotypic divergence in fruit and pistil traits (Fig. 1A and B). *De novo* genome assemblies were generated using 51.54 Gb PacBio HiFi data for each variety, achieving approximately 196× genome coverage (Tables S1 and S2). The initial assembly of S36 produced 2241 contigs with a contig N50 of 24.1 Mb, while the H19 assembly yielded 1946 contigs with a contig N50 of 22.0 Mb. Subsequent scaffolding was performed using 50.51 Gb of chromosome conformation capture sequencing (Hi-C) data (nearly 192× genome coverage), and chromatin interaction maps generated via

Table 1 Global characteristics of five cucumber genomes.

Assembly	North China type			South China type	
	S36	9930(v3)	XTMC	H19	Cu2
Size (Mb)	262.9	224.8	240.1	255.8	247.1
Year	2024	2019	2020	2024	2020
Number of chromosomes	7	7	7	7	7
Number of contigs	2241	174	926	1946	851
Contig N50 (Mb)	24.1	8.9	2.1	22	5.3
Number of genes	27 852	24 714	25 167	27 536	25 382
Number of gaps	6	86	175	4	91
Repeat content (%)	47.30	32.50	37.20	46.25	37.70
GC level (%)	35.50	32.56	32.51	35.09	32.53
BUSCOs (%)	98.8	95.5	96.6	98.7	97.6

Juicebox confirmed complete pseudomolecule organization. This process resulted in 1634 scaffolds for S36 (scaffold N50: 37.24 Mb) and 1433 scaffolds for H19 (scaffold N50: 34.60 Mb). Finally, 262.93 and 255.79 Mb of the assembled sequences were successfully anchored to the seven chromosomes of the S36 and H19 genomes, respectively (Fig. S1). Genome annotation identified 124.36 Mb (47.3%) of repetitive sequences in the S36 genome and 118.31 Mb (46.25%) in the H19 genome (Tables S3 and S4). Additionally, the genomes exhibit 35.50% and 35.09% GC content, respectively. Both repetitive sequence proportion and GC content are higher than those in other reported cucumber genomes, such as 9930 (v3), XTMC, and Cu2 genomes (Table S5) [8, 27]. Following repeat masking, 27 852 and 27 536 protein-coding genes were predicted in the S36 and H19 genomes, respectively, with average gene lengths of 3270 and 3311 bp, and 5.2 exons per gene on average (Table S6).

The quality and completeness of the S36 and H19 genome assemblies were assessed using multiple methods. First, 98.82% (1595 of 1614) and 98.76% (1594 of 1614) of conserved embryophyta BUSCO genes were identified in S36 and H19 genomes, respectively (Tables S7 and S8). Second, only six sequencing gaps were detected in S36 and four in H19, representing sequencing gaps between contigs (Tables S9 and S10)—fewer than those reported in previously published cucumber genomes (Table 1) [8, 27]. Third, 85.51% of genes in S36 and 86.10% in H19 were functionally annotated using integrated data from multiple databases, including NCBI nonredundant (NR), Gene Ontology (GO), SwissProt, InterProScan, The *Arabidopsis* Information Resource, and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Tables S11 and S12). These results collectively confirm the high completeness and annotation quality of both genome assemblies.

To elucidate structural divergence, a comparative synteny analysis identified 213 conserved homologous blocks between the S36 and H19 genomes (Fig. 1C). Notably, 221 unaligned regions spanning 15.25 Mb (3.18%) in S36 and 193 unaligned regions covering 8.12 Mb (5.90%) in H19 were detected (Table S13). Further investigation characterized gene distribution and transposable element (TE) dynamics across all seven chromosomes using 1-Mb windows (Fig. 1D).

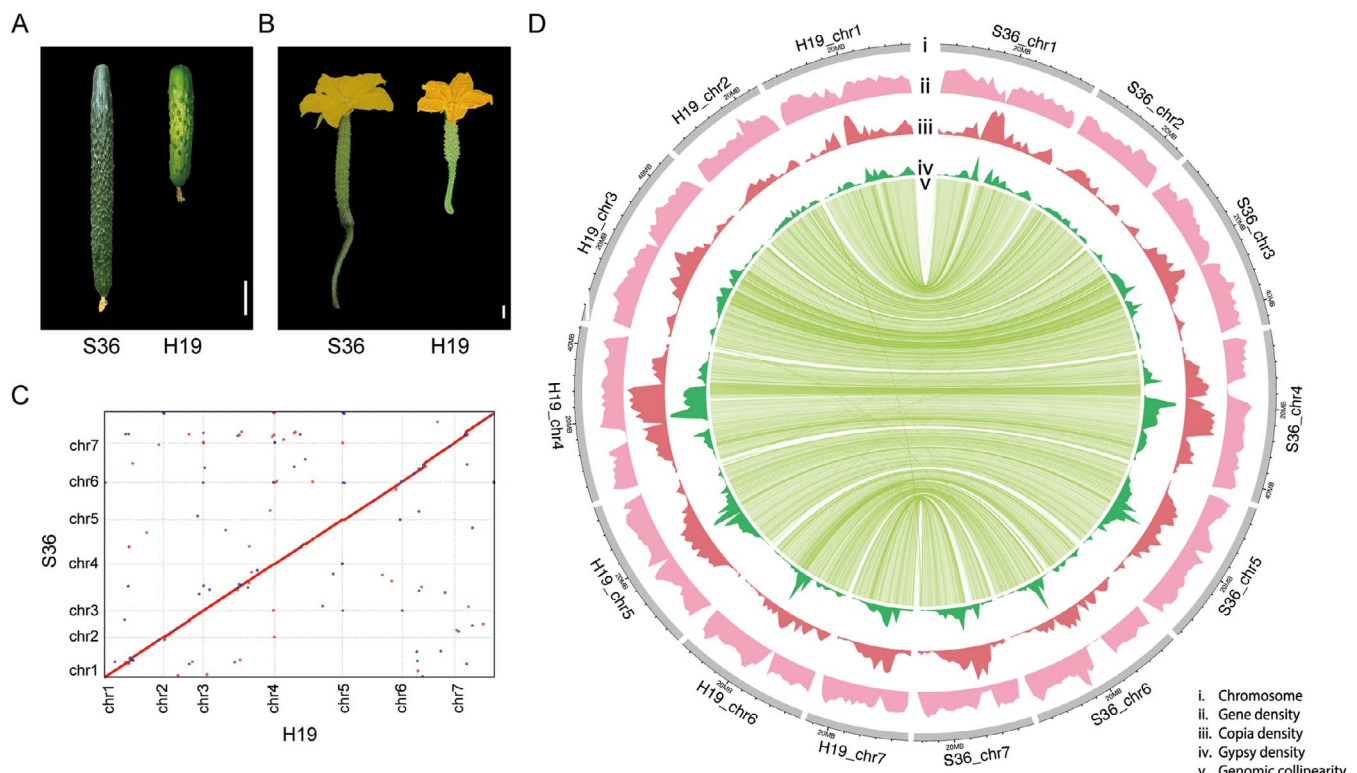


Figure 1 Phenotypic and genomic divergence between S36 and H19 genomes. (A) Fruit comparison between S36 (left) and H19 (right). Scale bar is 5 cm. (B) Pistil comparison between S36 (left) and H19 (right) on flowering day. Scale bar is 1 cm. (C) Syntenic collinearity analysis between S36 and H19 genomes. (D) Whole-genome architecture comparison between S36 and H19 genomes.

Structural variations between the North China type and South China type cucumber genomes

Structural variations play crucial roles during plant domestication, contributing to differences in the key characteristics [8]. To explore the genetic basis of these variations, the genomes of two North China type cucumber varieties (9930v3 and XTMC) and two South China type cucumber varieties (H19 and Cu2) were aligned to the S36 genome to identify genetic variants [8, 27]. After filtering out SVs smaller than 50 bp and those located in assembly gaps, 2391 SVs (approximately 22.45 Mb) were identified in the North China type genomes and 3765 SVs (approximately 20.28 Mb) in the South China type genomes. Additionally, 1700 SVs (~73.48 Mb) were specific to the North China type and 3074 (~61.62 Mb) were specific to the South China type (Fig. 2A). Most SVs were less than 5 kb in length (Fig. 2B), and the majority were located in non-TE regions (Fig. 2C). Moreover, these SVs significantly overlapped with 5219 genes in the S36 genome. Among these, 255 and 353 genes were specifically affected by breakpoints of North China type and South China type SVs, respectively. Notably, six of these genes showed high homology to the previously reported functional genes (Fig. 2D) [28–33]. In addition, polymerase chain reaction (PCR) validation of common type-specific SVs confirmed 91.9% of cases, further supporting the accuracy of SV detection (Fig. S5 and Table S14).

RNA-Seq analysis was performed using the RNA-seq data from the genomes of 9930 and H19, which revealed 7235 differentially expressed genes (DEGs). Of these, 64 and 101 DEGs overlapped with those affected by SV breakpoints in the North China type

and South China type genomes, respectively (Fig. S2). Functional enrichment analysis indicated that these genes primarily function in cellular processes, playing mainly intracellular and cytoplasmic roles in the North China type genomes, and a catalytic role in the South China type genomes (Figs S3 and S4). The sequence composition of SVs also differed markedly between the North China type and South China type genomes, particularly in terms of small RNA and DNA families (Fig. 2E).

Disease resistance in plants is often associated with SVs in the form of tandem arrays of resistance (R) genes [34, 35]. To identify R gene analogs between the previously reported North China type and South China type genomes, a homology analysis of these genes was conducted. A total of 881 R genes were identified in the S36 genome (Fig. S6), of which 42 (4.77%) overlapped significantly with those in the SVs. Among these genes, 14 and 22 were found to be affected by SVs specific to the North China type and South China type genomes, respectively. Additionally, the distribution of R genes was uneven across each chromosome (Fig. 2B). Subsequently, differential SNPs and variant genes were analyzed between the genomes. The South China type genomes were found to have more SNPs than the North China type genomes (928 493 versus 518 925), and the majority of these SNPs were classified as modified-effect ones (Tables S15–S17).

Comparison between the genomes of S36 and two North China type cucumber varieties

The S36 genome was compared with the 9930 (v3) and XTMC genomes to identify SVs [8, 27], revealing overall collinearity

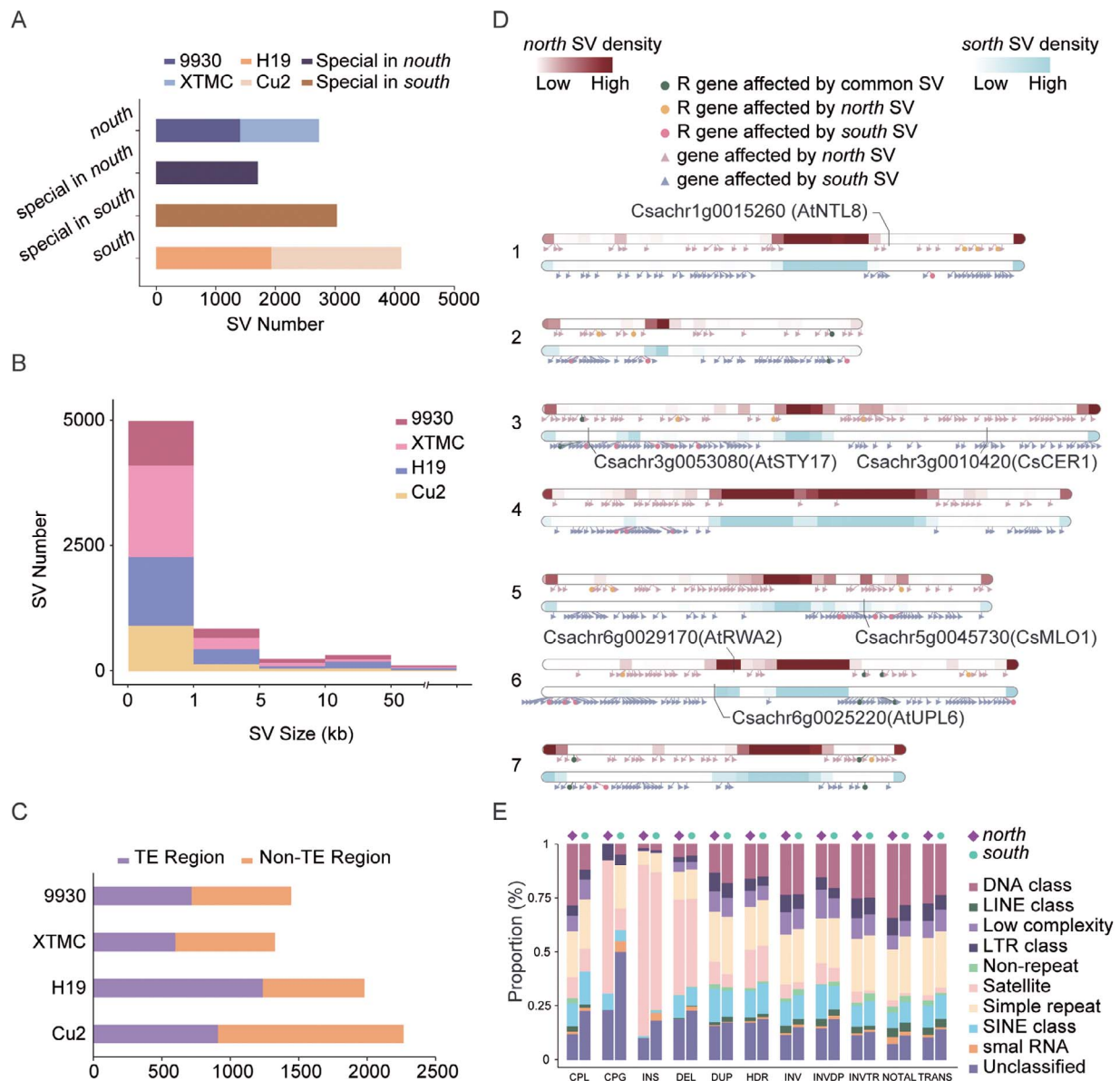


Figure 2 Structural variations between North China type and South China type genomes. 'north' represents North China type genomes. 'south' represents South China type genomes. (A) The number of structural variations between North China type and South China type genomes. (B) The length distribution of SVs among four cucumber genomes. (C) The number of SVs located in TE regions. (D) The distribution of SVs in the seven chromosomes and genes affected by North China type and South China type SVs, respectively. (E) The sequence composition of North China type and South China type SVs.

among the three genomes. Comparative genomic analysis identified 148 syntenic blocks between S36 and 9930, and 109 blocks between S36 and XTMC (Fig. 3A). Relative to the 9930 (v3) (210.94 Mb) and XTMC (204.50 Mb) assemblies, S36 included an additional 54.35 and 56.16 Mb of anchored sequences, respectively, predominantly localized to pericentromeric regions (Fig. S7). As a result, S36 chromosomes ranged from 26.45 to 46.14 Mb in length, exceeding those of 9930, which ranged from 22.47 to 40.88 Mb [27]. Next, the telomere-specific and centromere-specific repeats were screened in the genomes of S36 and the two North China type cucumber genomes. The results indicated the S36 genome possesses seven telomeres, more than the number of telomeres in any other genome (Table S18). The centromeres in all chromosomes of the S36 genome were significantly larger (Fig. 3B), indicating that the centromere assembly in the S36 genome is probably of higher quality. Long terminal

repeats (LTRs), the most abundant subgroup of TEs in cucumber, were also compared across the three genomes. The results showed that the number of newly inserted LTRs in the S36 genome is more than twice that in the other genomes (Fig. 3C). A further comparison of the genome components revealed that the proportion of TEs and total TE sequence length are significantly higher in the S36 genome (Fig. 3D). These data suggest that the S36 genome has undergone a stronger LTR expansion compared with the other cucumber genomes, possibly due to tissue culture [36, 37].

Whole-genome resequencing and screening of EMS populations

A mutant library of S36 was constructed through EMS mutagenesis. EMS treatment of cucumber seeds yielded 1785 M₂

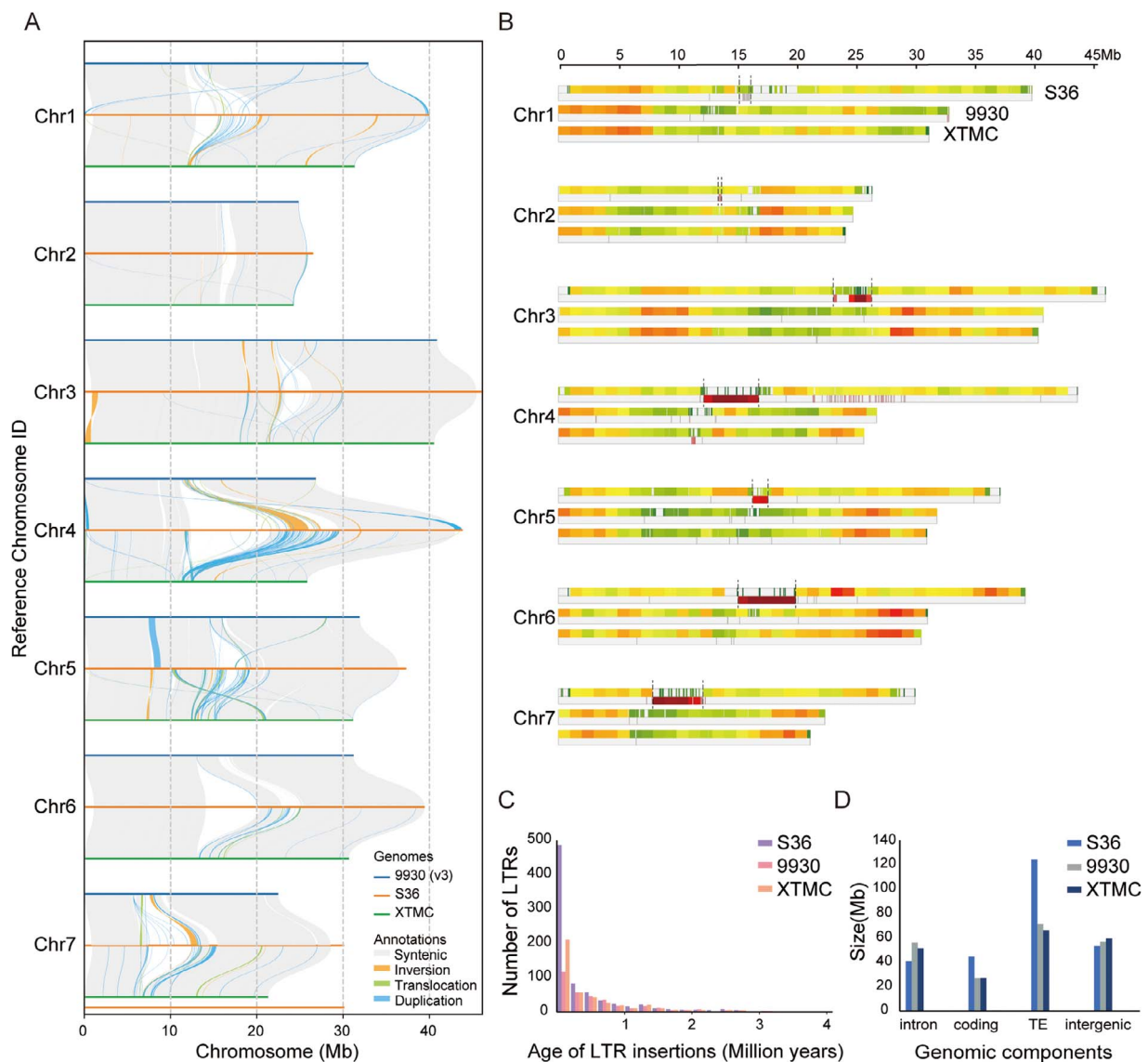


Figure 3 Comparison of S36 with the North China type cucumber genomes. (A) The collinearity map of the S36 and two North China type (9930 and XTMC) cucumber genomes. (B) Centromere identification results for three North China type (9930 and XTMC) cucumber genomes. Dashed lines denote centromeric regions. (C) Distributions of insertion times dated using intact LTRs in three North China type (S36, 9930, and XTMC) cucumber genomes. (D) Percentage and size of genomic components in three North China type (S36, 9930, and XTMC) cucumber genomes.

generation lines, of which 516 were cultivated, leading to the identification of 45 stably inherited mutants. These mutants exhibited notable differences from the wild-type in terms of various phenotypes, such as plant height, leaf shape and color, fruit length, internode length, and the density and size of warts (Fig. 4A–F). To further investigate the genetic basis of these phenotypic variations, whole-genome resequencing analysis was performed on nine representative mutants selected from the 45 identified mutants. The nine mutants yielded a total of 6.77 Gb of data, with the amount of data generated by an individual mutant ranging from 583.47 Mb (ES23) to 1.18 Gb (ES226), averaging 752.60 Mb, which corresponded to a sequencing depth of $29.24\times$ (Table S19). After screening and filtering the SNPs using a conventional approach, we excluded those in the TE regions from further analysis, retaining 80 536 SNPs distributed almost uniformly on each chromosome (Fig. 5E). Each mutant had an

average of 57 320 SNPs (Fig. 5A). On average, 5410 genes per mutant were affected by these SNPs (Fig. 5B). The number of SNPs per mutated gene varied greatly (Fig. 5C). In addition, the reference genome annotation revealed that 57.11% of 141 064 SNPs were located in the gene space, covering 17 318 genes (62.18% of all genes). These mutations were concentrated more in upstream regions (37.99%) than in downstream regions (19.64%) (Fig. 5D). Analyzing the SNP features of each mutant revealed the predominance of C/G to T/A and T/A to C/G mutations (30.04% and 29.88% on average, respectively) (Fig. 5F). The SnpEff program was used to predict the function of each mutated gene [38]. The majority of mutations (79.69% on average) were classified as the modified-effect ones; however, only a few of these (approximately 1.03%) were expected to have a strong effect (Table S20) and were located mainly in the upstream and downstream regions (Table S21).

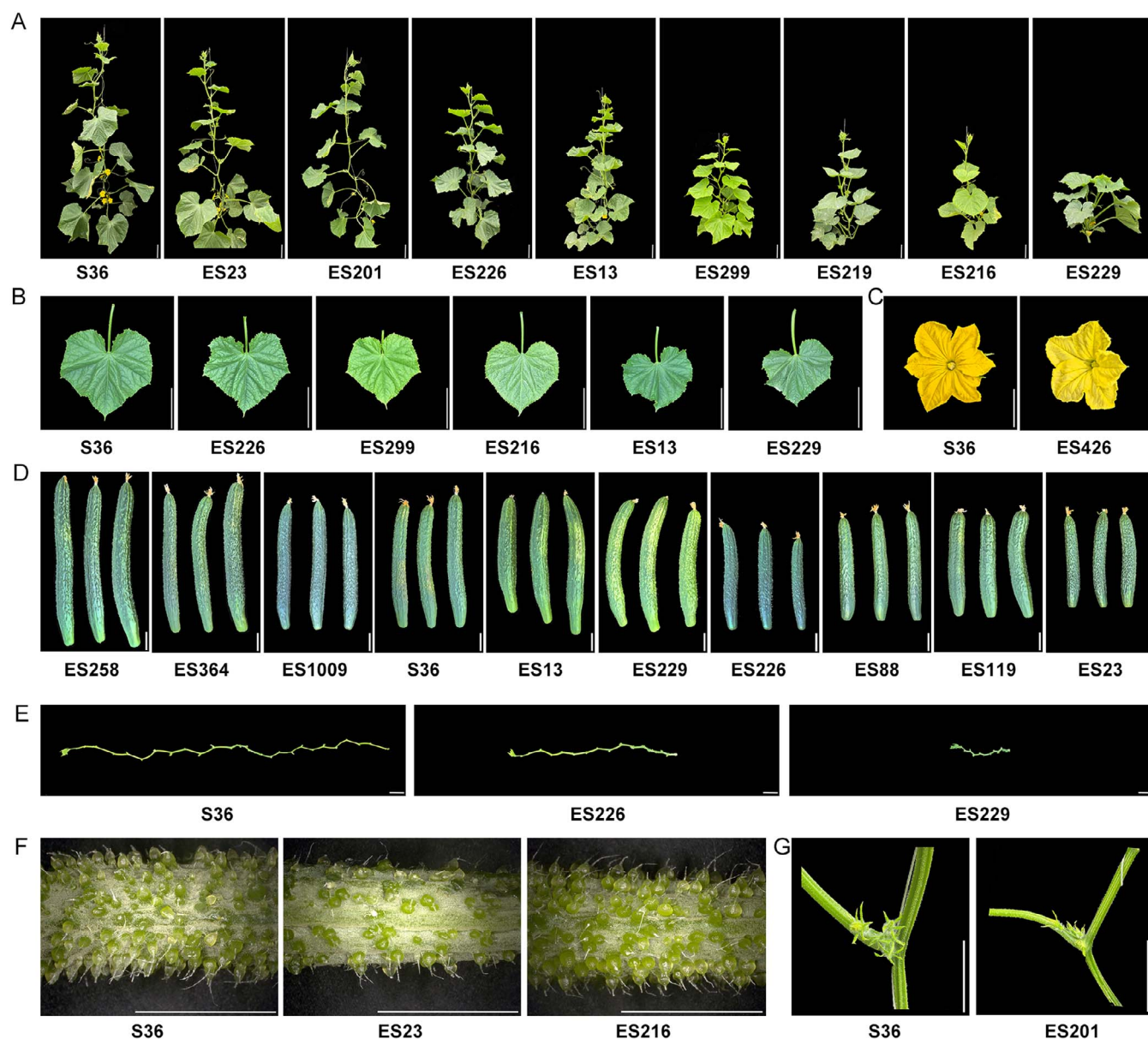


Figure 4 Phenotypic investigation of EMS mutant library. (A) Comparison of plant height between mutants and S36. Scale bar is 10 cm. (B) Leaf comparison between mutants and S36. Scale bar is 10 cm. (C) Comparison of flower colors. Scale bar is 2 cm. (D) Fruit comparison between mutants and S36. Scale bar is 5 cm. (E) Internode comparison between mutants and S36. Scale bar is 10 cm. (F) Magnified view of warts of mutants and S36. Scale bar is 1 cm. (G) Comparison of petiole-stem angle between mutant and S36. Scale bar is 10 cm. All cucumber materials were observed after three months of growth.

Fine mapping and functional verification of *CsCAO*

Leaf color mutants serve as a valuable reference for studying the genetic mechanisms underlying plant photosynthesis, chlorophyll biosynthesis, development, degradation, and tetrapyrrole synthesis, among others [39]. After 3 months of growth, phenotypic observations revealed that the mutant (ES299) exhibited yellow-green colored stems, petioles, ovaries, fruits, and leaves, which differed from the wild-type S36 (Fig. 6B). The total chlorophyll and carotenoid contents of leaves, petioles, stems, ovaries, and exocarps in ES299 were significantly lower than those in S36; however, the ratio of chlorophyll a to chlorophyll

b was relatively high in the mutant (Fig. 6C–E). Additionally, ES299 showed weaker growth potential than S36 during the same growth period, as evidenced by a noticeable decrease in its plant height, stem diameter, internode length, and leaf number (Fig. S8). Using the green leaf inbred line G35 and the yellow-green leaf ES299 mutant as parents in the hybridization experiment, the proportions of progenies exhibiting green and yellow-green leaves aligned with the expected segregation ratios of 3:1 and 1:1 in the F_2 and BC_2 populations, respectively, suggesting that the mutant trait was governed by a recessive gene (Table S22).

To further identify the candidate gene, four DNA pools (parental and progeny pools of G35 and ES299) were subjected to whole-genome resequencing. The high-quality reads were mapped to

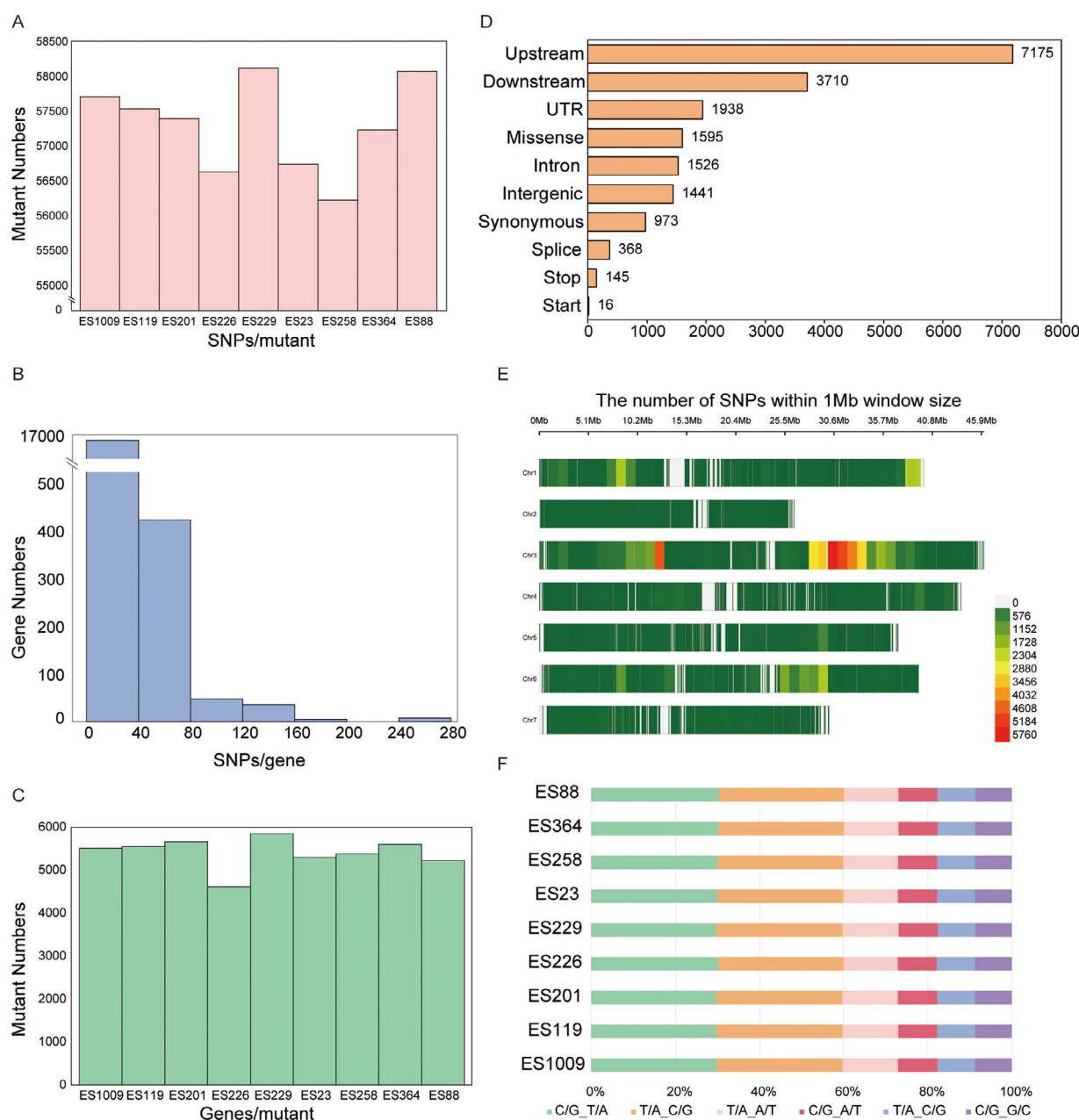


Figure 5 Characterization of the EMS-induced mutations in S36. (A) Distribution of SNPs in each mutant. (B) Distribution of the number of genes with mutations in each mutant. (C) Distribution of SNPs in each gene. (D) Distribution of SNPs with predicted effects on gene functions. (E) Mutation distribution and density for the mutations identified in mutants on seven chromosomes. (F) Ratios of different mutations identified in mutants.

the S36 genome and combined with the parental resequencing data. Preliminary mapping results for the yellow-green leaf trait were obtained through bulked segregant analysis (BSA) analysis using the Euclidean distance and SNP index association algorithm (Fig. 6A1). The confidence interval associated with leaf color traits was found on chromosome 6, which is located within the 9.5- to 18.1-Mb region. In this interval, 96 F_2 recessive plants were genotyped using 16 KASP markers (Fig. 6A2). Finally, the candidate region of the mutated trait was narrowed down to a 541-kb region flanked by the markers M13 and M16 (Fig. 6A3). This interval comprised five SNPs, four of which are located in the intergenic region, and one nonsynonymous SNP is located in the coding region of *Csa6G385090*; these SNPs resulted in an amino acid change from

leucine to phenylalanine at position 361 (Fig. 6A4 and Table S23). In the ES299 mutant, an L-to-F amino acid mutation occurs within the PobA domain of CsCAO, which is responsible for the catalytic activity of chlorophyllide a oxygenase. Three-dimensional structural analysis indicated that this mutation induced subtle changes in the local side-chain conformation (Fig. S9), potentially affecting substrate access to the catalytic pocket.

Csa6G385090 encodes a chlorophyll a oxygenase (CsCAO) that catalyzes the conversion of chlorophyll a into chlorophyll b. Three CsCAO-silenced lines (TRSV::CsCAO) with 70–80% reduction in CsCAO mRNA levels were obtained from 50 plants infected with TRSV-CsCAO by using the TRSV-VIGS system (Fig. 6F). The leaves, petioles, and stems of the TRSV::CsCAO lines exhibited a lighter

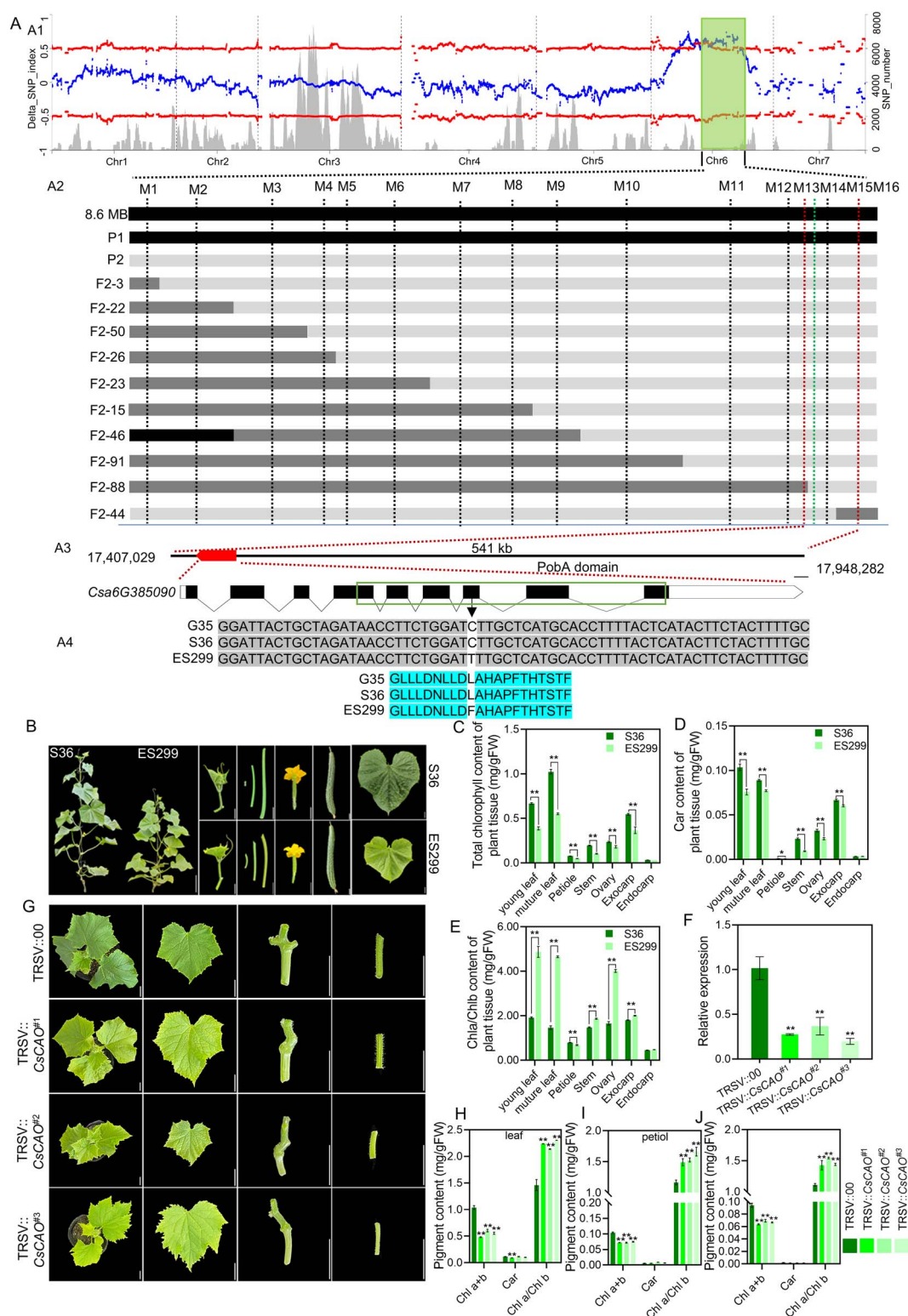


Figure 6 Gene mapping of ES299 mutant and functional verification of *CsCAO*. (A1) Δ(SNP index) of all cucumber chromosomes. SNP index peak is found between 9.5 and 18.1 Mb on chromosome 6. (A2) The genotyping of recombinant plants from F₂ population using the 16 markers allowed mapping the mutant gene in a 541-kb region of chromosome 6. (A3) The location and structure of *Csa6G385090* in the region of 541 kb. The white boxes indicate the 5'UTR and 3'UTR positions of *Csa6G385090*, and the black boxes and broken lines represent the positions where exons and introns are located, respectively. (A4) Nucleotide and protein sequence alignment of *Csa6G385090* in G35, S36, and ES299. (B) Phenotypic observation of S36 and ES299. From left to right: whole plant (scale bar, 20 cm), stem, petiole, ovary, fruit, and leaf (scale bar, 3 cm). (C–E) The total chlorophyll, carotenoid, and the ratio of chlorophyll a to chlorophyll b in different tissues of S36 and ES299 are determined. Values are means ± SE (n = 4). (F) Identification of *TRSV::CsCAO* silencing lines by qRT-PCR. Values are means ± SE (n = 3). (G) Phenotypic observation of *TRSV::00* and *TRSV::CsCAO*. From left to right: whole plant, leaf, stem, and petiole (scale bar, 2 cm). (H–J) The total chlorophyll, carotenoid, and the ratio of chlorophyll a to chlorophyll b in different tissues of *TRSV::00* and *TRSV::CsCAO* are determined. Values are means ± SE (n = 3). Student's *t* test is used to test the significance of the data. **P* < 0.05; ***P* < 0.01.

coloration than those of the control lines (TRSV::00) (Fig. 6G). Furthermore, the TRSV::CsCAO lines exhibited significantly lower contents of chlorophyll in the leaves, petioles, and stems but a higher ratio of chlorophyll a to chlorophyll b (Fig. 6H–J). In summary, the phenotypic traits of the CsCAO-silenced plants were similar to those of the ES299 mutant, highlighting that CsCAO silencing inhibited chlorophyll synthesis and resulted in lighter color of the organs in cucumber. These results reinforce the utility of the EMS mutant library construction for identifying candidate genes associated with visible phenotypic traits using forward genetic approaches [15]. Furthermore, these results validate the reliability of genome assemblies combined with mutant libraries for a rapid functional gene identification, providing a reliable pathway for the functional characterization of genes.

Discussion

Historically, cucumber is believed to have originated in the Himalayan Mountains, with its domestication dating back to nearly 3500 years ago [23]. According to morphological characteristics and geographical distribution, it is classified into four types as follows: Indian, Eurasian, East Asian, and Xishuangbanna [23, 24]. Similarly, Chinese cucumber varieties have been classified into four subgroups—South China type, North China type, Xishuangbanna type, and Europe type; of these, the first two types have originated from the previously reported East Asian group [24, 40]. Although molecular markers have been commonly used to investigate the genetic diversity of cucumber germplasm resources in China, only a few markers or varieties have been tested to date, and a comprehensive study of cultivated Chinese varieties is lacking [24, 40, 41]. Although North China and South China type cucumber varieties exhibit significant differences in fruit length, wart size and density, and other phenotypic traits, a comparative genomic analysis between North and South types based on genomic sequences has not been reported yet.

The South China and North China types are unique cucumber varieties in China, and the market demand and consumption of both varieties remain high [25]. S36 and H19 cucumber varieties have become core germplasm in commercial cucumber breeding. Therefore, the high-quality genome assembly of S36 and H19 can serve as a valuable resource for the identification of candidate genes related to vital agronomic traits in cucumber. In this study, we assembled S36 and H19 genomes, both of which have higher quality than those of previously published cucumber genomes [8, 27]. However, additional research is needed to construct a gap-free assembly of the cucumber genome, which could not be achieved in this study. In this study, the representative genomes of the two cucumber ecotypes were selected to identify the SVs specific to each type and the genes they affect, providing insights into the classification and evolutionary pattern of cucumber varieties. Among these, *Csachr3g0053080* (*AtRWA2*) and *Csachr3g0053080* (*AtSTY17*), associated with resistance to *Botrytis cinerea*, and *Csachr5g0045730* (*CsMLO1*), associated with resistance to powdery mildew disease, were found to be affected by north- and south-specific SVs. These findings suggest that the three genes may be related to the disease susceptibility of South China type cucumber varieties [28, 32, 33]. Similarly, *Csachr1g0015260* and *Csachr6g0025220*, which are

highly homologous to *AtNTL8* and *AtUPL6*, respectively, were found to regulate trichome formation in *Arabidopsis* and may be related to the differences in terms of fruit wart phenotype between the South and North varieties [29, 30]. Previous studies have demonstrated that waxy fruits exhibit lower surface gloss [31]. *Csachr3g0010420* (*CsCER1*), which is involved in wax metabolism, was found to influence fruit skin glossiness and may contribute to the difference in glossiness between the South and North types [31, 42]. The discovery of these genes lays a foundation for the classification of cucumber varieties and further research on the genetic factors affecting the phenotypic differences among them. Moreover, differences in the TE composition of SVs between the genomes of North and South types suggested that TEs might have been subjected to differential selection due to differences in the environmental condition between the north and south regions. Furthermore, significant differences were observed in the chromosomal distribution and number of R genes affected by SVs, suggesting that these genes might have undergone differential selection due to differences in individuals' dietary habits between the northern and southern regions [24]. While the identified SVs serve as a valuable resource for cucumber breeding, further functional validation of these SVs is necessary to understand their precise role in phenotypic trait regulation. The expression of these SVs might have also been influenced by the genetic background of different cucumber varieties, warranting further investigation in this direction. Moreover, in addition to the North and South China type cucumber varieties, other cucumber varieties exist, which contain favorable genes. Introgression of these genes in crop breeding can enrich the genetic diversity of the South China and North China type cucumber [40].

This study considered S36 as the reference genome and performed whole-genome resequencing on nine EMS mutants. The analysis yielded systematic phenotypic data of the mutants, along with the associated genomic information, which can advance functional genomics research in cucumber. Additionally, the mutants generated in this study can be used directly in cucumber breeding programs to introduce favorable traits, such as high flowering rate and pathogen resistance. This study also identified and annotated various TEs, including newly inserted, intact LTRs, in the S36 genome. Moreover, the nine mutant lines were selected for sequencing analysis, which enabled the accurate and efficient detection of mutations. Through whole-genome sequencing of these nine mutants, we identified 80 536 SNPs, located in both coding and noncoding regions, which can serve as a valuable resource for future functional studies [43].

A high-quality genome, combined with a large corresponding mutant library, facilitates the identification and cloning of candidate genes [44, 45]. This strategy has been used successfully in watermelon [16] and Chinese cabbage [15] to clone many genes from EMS libraries. In plants, chlorophyll is an essential pigment that plays a crucial role in energy transfer and transformation in photoreactions by absorbing solar energy and binding to various Chl-binding proteins [46]. In this study, we utilized the yellow-green leaf mutant ES299 to investigate the genetic basis of chlorophyll biosynthesis. Leveraging the S36 genome and through whole-genome resequencing of 9 mutant plants, we identified CsCAO as a candidate gene for chlorophyll synthesis from the EMS mutant library. CsCAO mutants demonstrated disrupted chlorophyll biosynthesis and thus a lighter coloration of their organs. In summary, the new reference genome, together with

the EMS mutant library, serves as a powerful tool for the genetic analysis and enhancement of cucumber traits, as well as a reliable pathway for determining gene functions.

Through the genome assembly and mutation analysis of South China and North China type cucumber varieties, the present study sheds light on the classification and evolution of Chinese cucumber varieties. The high-quality assembly and annotation of the S36 genome enhanced the accuracy and efficiency of mutation detection, underscoring the usefulness of the forward genetic approach as a valuable tool for functional genomics research in cucumber. Using this approach to identify genes associated with desirable agronomic traits can expedite advancements in cucumber breeding. Finally, the variants identified in this study can facilitate improvements in key agronomic traits of not only cucumber but also other closely related Cucurbitaceae crops.

Materials and methods

Plant materials and sequencing

The S36 cucumber variety was derived from the commercial cultivar ‘Ke Run 99’ through 10 generations of continuous self-pollination. Two commercial varieties, ‘Wei Lai 103’ and ‘Jin Mei Han Yu’, were crossed, and the hybrid offspring were self-pollinated for eight generations to obtain the stable material H19. Both were planted in the Tianjin Academy of Agricultural Sciences (TAAS). The leaves of S36 and H19 were used to construct the PacBio HiFi library for genome sequencing [47]. The Hi-C libraries were built according to the Proximo Hi-C plant protocol with the restriction enzyme MboI [48]. The sequencing libraries were sequenced using the PacBio Sequel II/Ile sequencing platform or the Revio platform at Berrygenomics Company. Optical equipment was used to convert the raw data into the initial output file, Polymerase reads. These Polymerase reads were then subjected to basic filtering using the instrument’s built-in software, SMRT Link, and were subsequently converted into Subreads.bam (Sequel II/Ile CLR/CCS mode), Reads.bam (Sequel IIe CCS mode), or hifi_read.bam (Revio sequencing).

De novo assembly of S36 and H19 genome

Hifiasm, an efficient open-source *de novo* assembler specifically designed for HiFi reads, was used to extract overlaps and build the assembly graph. This approach enabled the separation of distinct alleles or different copies of segmental duplications containing a single segregating site [49]. A preliminary contig reference genome of approximately 387 Mb for S36 and 372 Mb for H19 was obtained, and the genome continuity was evaluated based on the contig N50 length. Hi-C reads were aligned to the enhanced contigs using Juicer (V1.5) for feature analysis and data extraction [50]. The output was processed with 3D-DNA to correct misjoins, as well as to order, orient, and scaffold the sequences, resulting in an improved assembly [51]. Finally, Juicebox was used to visualize and interactively assemble the genome by manually adjusting chromosome boundaries and fixing some minor errors [52]. The completeness of the assembled genomes was evaluated using BUSCO (V5.2.1) with the embryophyta_odb10 database [53, 54]. Synteny analysis was conducted using the MUMmer package (V3.23) to compare the assembled genomes with the 9930 (v3) genome [55]. First,

NUCmer was used to perform comparisons between genomes with the parameters -mum -mincluster 100. Subsequently, Delta-filter was used to filter the alignment file generated by NUCmer with the parameters -l 1000 -l 1. Finally, a dotplot was generated with mummerplot [56], and the chromosomes of S36 and H19 were renamed according to the sequence of 9930 chromosomes. Furthermore, SYRI (V1.5.4) [57] was used to perform a genome-wide comparison of SVs between the assembled genomes and the 9930 genome.

Structural annotation and functional annotation of genes

The genome repeats were identified and annotated using RepeatModeler (V2.0.1) [58] and RepeatMasker (V4.1.0) [59] based on a custom repeat sequence library. Gene prediction was performed on the masked genome sequences using three complementary approaches: homologous prediction, RNA-seq-based prediction, and *de novo* prediction [15]. For RNA-seq-based prediction, transcriptomic data were generated from mixed samples of S36 and H19 (including leaves, female flowers, male flowers, fruit, and tendrils), 9 near-isogenic line materials (pericarp), and 10 different stages and sites of 9930 (including leaves, tendrils, roots, ovaries, female flowers, male flowers, and stems) [60]. These data were analyzed using Trinity and PASA to predict genes [61, 62]. For homologous prediction, protein sequences from 10 cucurbitaceae plants were retrieved from the Cucurbit Genomics Database (<http://cucurbitgenomics.org>), and ProtHint was used to align and splice predicted genes to a reference protein database [63]. For *de novo* prediction, the processed transcriptomic and protein data were mapped to the S36 genome, and protein-coding genes were predicted using AUGUSTUS [64] and MAKER [65]. The predicted annotations were validated using in-house Python scripts to ensure the correct placement of start and stop codons, and genes containing internal stop codons were removed. Finally, genes with coding sequences (CDSs) shorter than 300 bases were filtered out to obtain the final gene structural annotation. The functions of the S36 and H19 protein-coding genes were predicted using NCBI (NR), SwissProt, and the *Arabidopsis* database via Diamond (V0.9.24.125) [66]. Additionally, protein domain and gene ontology term annotations were performed using InterProScan (V5.59) [67], along with the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (<http://egg-nog-mapper.embl.de/>) Automatic Annotation Server [68]. DRAGO2, a tool of PRGdb (V4.0), was used to predict R genes in the S36 genome [69].

Genome-wide comparison of cucumber genomes

The genotypic data of our 210 materials were filtered for 4D SNPs [70] using a custom script. These 4-fold degenerate sites were then used for phylogenetic analysis with PHYLIP [71], and the phylogenetic tree was visualized using iTOL (<https://itol.embl.de>). Genome data for two North China type cucumbers (9930, XTMC) and one South China type cucumber (Cu2) were retrieved from Cucumber-DB (<http://www.cucumbersdb.com>) and CuGenDB (<http://cucurbitgenomics.org>). Genome-wide comparisons between the S36 cucumber genome and other

cucumber genomes were conducted using minimap2 [72] and the MUMmer package (V3.23) [55]. SVs between seven cucumber genomes were identified using SYRI(V1.5.4) [57]. SnpEff was used to annotate SVs larger than 50 bp [38]. A custom script was employed to calculate the sequence types and repeat sequence content associated with the SVs. Genes within the SV regions were considered potentially impacted, and the distribution of SVs, along with the functional gene and R genes affected by these SVs, was visualized using the Rldeogram package [73]. To construct LTR libraries, LTR_Finder (V1.07) [74] was used with default settings, and LTR_harvest (V1.6.1) [75] was employed with parameters '-minlenltr 100 -maxlenltr 7000 -motif TGCA -similar 90 -seed 20'. The results from LTR_Finder and LTR_harvest were merged using LTR_retriever (V2.9.0) [76] to generate the final LTR library. Subsequently, the LTR library was combined with the TE library, and LTR insertion times were calculated using a custom script. RepeatModeler (V2.0.1) [58] and RepeatMasker [59] were used to annotate and classify based on the constructed library with default parameters. The centromeres of the genome were determined using TandemRepeatFinder (TRF) [77], telomeres were identified using quarTeT [78], and the density of genes and centromeres was calculated using a custom script. These results were then visualized using Rldeogram [73].

EMS treatment and phenotypic investigation

Over 10 000 seeds with a 98% germination rate from full and plump grains were selected and soaked in double-distilled water for 4 h. The seeds were then treated with a 0.4% (W/V) EMS solution and shaken on a shaker for 12 h (EMS purchased from Sigma). Subsequently, the seeds were immersed in a 5% Na₂S₂O₃ (sodium thiosulfate) solution for 2 h (detoxification), washed with tap water for 2 h, and placed in a temperature-controlled box for germination. In early spring, the seedlings were grown in a greenhouse with regular management in TAAS. Then, they were transplanted into the greenhouse, and phenotypic data were recorded regularly. During flowering, female flowers with more than 15 nodes were selected for strict self-pollination. After fruit maturation (approximately 40 days postpollination), seeds were collected from individual plants, labeled, and used for summer sowing. Many mutants appeared in this generation, and a phenotypic survey was conducted, followed by strict self-pollination. Seeds were then collected from individual plants, and mutant seeds from individual plants were planted to observe and measure phenotypes.

Mutation detection in the mutants

The clean reads of mutants were aligned to the S36 genome using the BWA-MEM (V0.7.71) [79] with default parameters, and mapping results were obtained in SAM format. These results were subsequently processed using SAMtools (V1.9) [80] to convert the SAM format to BAM format, sort the BAM file, and obtain a consensus genotype for each locus. BCFtools (V1.18) [81] was then employed to convert the BAM format to VCF format. High-quality SNPs and INDELs (QUAL ≥ 30, DP ≥ 2, MQ ≥ 30) were used for subsequent mutation analysis. Additionally, SVs in the mutants were predicted using DELLY (V0.8.7) [82].

Mutant materials and growth conditions

A yellow-green leaf mutant, ES299, was obtained by EMS mutagenesis of the green leaf inbred line S36. Genetic analysis and gene mapping were carried out using distant green leaf advanced inbred line G35 and mutant ES299 as parents. All of the above cucumber plants were grown in a plastic greenhouse at 28–32°C under the natural light conditions provided by the Tianjin Kerun Cucumber Research Institute. The VIGS experiment was carried out using North China type cucumber XTMC as the material. The infected seedlings were cultivated in an artificial climate chamber with a 16-h light (22°C) and 8-h dark (18°C) cycle.

Determination of growth indexes

The vertical distance from the cotyledon node to the apical bud (plant height) and the distance between two adjacent nodes (node length) of wild-type S36 and mutant ES299 grown for 3 months were measured with a ruler. The stem diameter of the cotyledon node position was measured with a Vernier caliper. The total number of leaves of plants growing for 3 months was counted. Each index was measured for at least six biological replicates.

Determination of pigment contents

The pigment contents were determined according to the standard method of Lichtenthaler [83]. Cucumber tissues (0.2 g) were chopped and extracted with 95% ethanol for 24 to 48 h until the samples no longer faded. Subsequently, the absorbance of the extracts was measured at 665, 649, and 470 nm by a microplate fluorometer, with each measurement repeated three times [84]. According to the formula, the contents of chlorophyll a, chlorophyll b, and carotenoids were calculated [85].

Genetic mapping of candidate genes

G35 was crossed with ES299 to produce F₁ generation, and F₁ plants were selfed and backcrossed to obtain F₂ and BC₂ populations, respectively. Subsequently, the number of green leaf plants and yellow-green leaf plants in the progeny segregating population was counted, and the Chi-square (χ^2) test was used to analyze the trait separation rate. The equal amounts of DNA from green leaf and yellow-green leaf extreme phenotypes in F₂ population were selected to construct two extreme sequencing mixed pools, and parental DNA was used to construct the parental pools for sequencing analysis. The sequencing reads were aligned to the cucumber reference genome (9930v3). Subsequently, SAMtools [80] and GATK [86] software were used to process the data to obtain high-quality SNPs. All SNPs were annotated and mapped to seven chromosomes of cucumber. The SNP index and Δ SNP index were calculated to determine the chromosomal regions linked to the mutant phenotype and possible mutation sites [87]. Based on the analysis results of MutMap, KASP genotyping was performed on the candidate SNPs to further determine the candidate genes. A total of 96 F₂ plants were used for KASP genotyping. The KASP thermal cycle conditions are programmed according to the description of Xi *et al.* (2018) [88].

Tobacco ringspot virus-base-VIGS system in cucumber

The specific CDS fragment of CsCAO was inserted into pTRSV2 vector and then transformed into *Agrobacterium tumefaciens* GV3101. The *Agrobacterium* solutions containing the pTRSV1 and pTRSV2 vectors were mixed in equal volumes and incubated for 3 h at 28°C. The mixed bacterial solution was infected with XTMC cucumber seed buds for 20 min under a vacuum condition of −900 kPa. Subsequently, the seeds were placed at 25°C in the dark for 3 days. Finally, the cocultured seeds were planted in a light incubator (22°C/16-h light, 18°C/8-h dark) for a 1-month culture. The expression level of CsCAO was detected by quantitative real-time polymerase chain reaction (qRT-PCR), and plants with a 60% to 80% decrease in CsCAO expression were selected for subsequent experiments [89, 90].

Conserved domain prediction and three-dimensional structure analysis

The conserved domains of the CsCAO protein were predicted using the NCBI Conserved Domain Database (CDD) search tool (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>). The three-dimensional structure of CsCAO was modeled and downloaded in PDB format. Structural visualization and comparative analysis of the wild-type and mutant proteins were performed using PyMOL, with a focus on alterations in secondary structure, spatial conformation, and potential active sites.

Acknowledgments

The research was supported by the State Key Laboratory of Vegetable Biobreeding (SKLVB202506 to X.L.), the National Key Research and Development Program of China (2023YFF1000100 to T.L.), the Tianjin Major Project for Seed Industry (23ZXZYSN00010), the 111 Project (B17043 to T.L.), and the Construction of Beijing Science and Technology Innovation and Service Capacity in Top Subjects (CEFF-PXM2019_014207_000032 to T.L.).

Author Contributions

H.H. and T.L. conceived and designed the study. H.H., J.W., W.K., W.M., L.Z., Y.L., and J.L. planted and prepared the materials. J.H. performed the bioinformatics analysis. J.W. and X.L. designed and performed molecular experiments. J.H. and J.W. wrote the manuscript. T.L., H.H., and X.L. edited and improved the manuscript. All authors approved the final manuscript.

Data availability

The original sequencing data for genome assembly and annotation as well as the sequencing data from mutants and bulked pools for BSA analysis have been deposited in the BIG Submission Portal (<https://ngdc.cncb.ac.cn/>) under BioProject accession number PRJCA038364. The assembled genomes S36 and H19 were deposited in the Genome Warehouse (GWH) database of the Big Data Center (<https://bigd.big.ac.cn/gwh/>) under the accession number GWHGDGZ00000000.1 and GWHGDGY00000000.1, respectively.

Conflicts of interest statement

The authors declare no competing interest.

Supplementary material

Supplementary material is available at *Horticulture Research* online.

References

1. Guan J, Miao H, Zhang Z. *et al.* A near-complete cucumber reference genome assembly and Cucumber-DB, a multi-omics database. *Mol Plant*. 2024;17:1178–82
2. Lin T, Zhu G, Zhang J. *et al.* Genomic analyses provide insights into the history of tomato breeding. *Nat Genet*. 2014;46:1220–6
3. Zhao G, Lian Q, Zhang Z. *et al.* A comprehensive genome variation map of melon identifies multiple domestication events and loci influencing agronomic traits. *Nat Genet*. 2019;51:1607–15
4. Xie D, Xu Y, Wang J. *et al.* The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. *Nat Commun*. 2019;10:5158
5. Lyu X, Xia Y, Wang C. *et al.* Pan-genome analysis sheds light on structural variation-based dissection of agronomic traits in melon crops. *Plant Physiol*. 2023;193:1330–48
6. Liu Y, du H, Li P. *et al.* Pan-genome of wild and cultivated soybeans. *Cell*. 2020;182:162–176.e13
7. Cai X, Chang L, Zhang T. *et al.* Impacts of allopolyploidization and structural variation on intraspecific diversification in *Brassica rapa*. *Genome Biol*. 2021;22:166
8. Li H, Wang S, Chai S. *et al.* Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat Commun*. 2022;13:682
9. Della Coletta R, Qiu Y, Ou S. *et al.* How the pan-genome is changing crop genomics and improvement. *Genome Biol*. 2021;22:3
10. Jayakodi M, Schreiber M, Stein N. *et al.* Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res*. 2021;28:dsaa030
11. Bayer PE, Golicz AA, Scheben A. *et al.* Plant pan-genomes are the new reference. *Nat Plants*. 2020;6:914–20
12. Hübner S. Are we there yet? Driving the road to evolutionary graph-pangenomics. *Curr Opin Plant Biol*. 2022;66:102195
13. Shang L, Li X, He H. *et al.* A super pan-genomic landscape of rice. *Cell Res*. 2022;32:878–96
14. Tran QH, Shang L, Kappel C. *et al.* Mapping-by-sequencing via MutMap identifies a mutation in ZmCLE7 underlying fasciation in a newly developed EMS mutant population in an elite tropical maize inbred. *Genes (Basel)*. 2020;11:281
15. Sun X, Li X, Lu Y. *et al.* Construction of a high-density mutant population of Chinese cabbage facilitates the genetic dissection of agronomic traits. *Mol Plant*. 2022;15:913–24
16. Deng Y, Liu S, Zhang Y. *et al.* A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol Plant*. 2022;15:1268–84

17. Fu A, Zheng Y, Guo J. *et al.* Telomere-to-telomere genome assembly of bitter melon (*Momordica charantia* L. var. *abbreviata* Ser.) reveals fruit development, composition and ripening genetic characteristics. *Hortic Res.* 2023;10:uhac228
18. Jiang F, Guo M, Yang F. *et al.* Mutations in an AP2 transcription factor-like gene affect internode length and leaf shape in maize. *PLoS One.* 2012;7:e37040
19. Zhao J, Jiang L, Che G. *et al.* A functional allele of CsFUL1 regulates fruit length through repressing CsSUP and inhibiting auxin transport in cucumber. *Plant Cell.* 2019;31:1289–307
20. Niu H, Liu X, Tong C. *et al.* The WUSCHEL-related homeobox1 gene of cucumber regulates reproductive organ development. *J Exp Bot.* 2018;69:5373–87
21. Cheng F, Song M, Zhang M. *et al.* A SNP mutation in the CsCLAVATA1 leads to pleiotropic variation in plant architecture and fruit morphogenesis in cucumber (*Cucumis sativus* L.). *Plant Sci.* 2022;323:111397
22. Lai W, Zhou Y, Pan R. *et al.* Identification and expression analysis of stress-associated proteins (SAPs) containing A20/AN1 zinc finger in cucumber. *Plants (Basel).* 2020;9:400
23. Renner SS, Schaefer H, Kocyan A. Phylogenetics of *Cucumis* (Cucurbitaceae): cucumber (*C. sativus*) belongs in an Asian/Australian clade far from melon (*C. melo*). *BMC Evol Biol.* 2007;7:58
24. Lv J, Qi J, Shi Q. *et al.* Genetic diversity and population structure of cucumber (*Cucumis sativus* L.). *PLoS One.* 2012;7:e46919
25. Yu B, Ming F, Liang Y. *et al.* Heat stress resistance mechanisms of two cucumber varieties from different regions. *Int J Mol Sci.* 2022;23:1817
26. Huang H, Yang Q, Zhang L. *et al.* Genome-wide association analysis reveals a novel QTL CsPC1 for pericarp color in cucumber. *BMC Genomics.* 2022;23:383
27. Li Q, Li H, Huang W. *et al.* A chromosome-scale genome assembly of cucumber (*Cucumis sativus* L.). *GigaScience.* 2019;8:giz072
28. Dong S, Liu X, Han J. *et al.* CsMLO8/11 are required for full susceptibility of cucumber stem to powdery mildew and interact with CsCRK2 and CsRbohD. *Hortic Res.* 2023;11:uhad295
29. Tian H, Wang X, Guo H. *et al.* NTL8 regulates trichome formation in *Arabidopsis* by directly activating R3 MYB genes TRY and TCL11. *Plant Physiol.* 2017;174:2363–75
30. Downes BP, Stupar RM, Gingerich DJ. *et al.* The HECT ubiquitin-protein ligase (UPL) family in *Arabidopsis*: UPL3 has a specific role in trichome development. *Plant J.* 2003;35:729–42
31. Wang W, Zhang Y, Xu C. *et al.* Cucumber ECERIFERUM1 (CsCER1), which influences the cuticle properties and drought tolerance of cucumber, plays a key role in VLC alkanes biosynthesis. *Plant Mol Biol.* 2015;87:219–33
32. Manabe Y, Nafisi M, Verherthbruggen Y. *et al.* Loss-of-function mutation of REDUCED WALL ACETYLATION2 in *Arabidopsis* leads to reduced cell wall acetylation and increased resistance to *Botrytis cinerea*. *Plant Physiol.* 2011;155:1068–78
33. Chen L, Xiao J, Li Y. *et al.* The Raf-like MAPKKKs STY8, STY17, and STY46 negatively regulate *Botrytis cinerea* resistance by limiting MKK7 protein accumulation in *Arabidopsis*. *Plant J.* 2024;117:1503–16
34. Chovelon V, Feriche-Linares R, Barreau G. *et al.* Building a cluster of NLR genes conferring resistance to pests and pathogens: the story of the Vat gene cluster in cucurbits. *Hortic Res.* 2021;8:72
35. McHale LK, Haun WJ, Xu WW. *et al.* Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* 2012;159:1295–308
36. Hirochika H, Sugimoto K, Otsuki Y. *et al.* Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Natl Acad Sci U S A.* 1996;93:7783–8
37. Komatsu M, Shimamoto K, Kyoizuka J. Two-step regulation and continuous retrotransposition of the rice LINE-type retrotransposon karma. *Plant Cell.* 2003;15:1934–44
38. Cingolani P, Platts A, Wang LL. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, *SnpEff. Fly (Austin).* 2012;6:80–92
39. Chen N, Wang P, Li C. *et al.* A single nucleotide mutation of the *IspE* gene participating in the MEP pathway for isoprenoid biosynthesis causes a green-revertible yellow leaf phenotype in rice. *Plant Cell Physiol.* 2018;59:1905–17
40. Zhang J, Yang J, Zhang L. *et al.* A new SNP genotyping technology Target SNP-seq and its application in genetic analysis of cucumber varieties. *Sci Rep.* 2020;10:5623
41. Qi J, Liu X, Shen D. *et al.* A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat Genet.* 2013;45:1510–5
42. Huang H, du Y, Long Z. *et al.* Fine mapping of a novel QTL CsFSG1 for fruit skin gloss in cucumber (*Cucumis sativus* L.). *Mol Breed.* 2022;42:25
43. LeVan TD, Guerra S, Klimecki W. *et al.* The impact of CD14 polymorphisms on the development of soluble CD14 levels during infancy. *Genes Immun.* 2006;7:77–80
44. Nie S, Wang B, Ding H. *et al.* Genome assembly of the Chinese maize elite inbred line RP125 and its EMS mutant collection provide new resources for maize genetics research and crop improvement. *Plant J.* 2021;108:40–54
45. Tang S, Liu DX, Lu S. *et al.* Development and screening of EMS mutants with altered seed oil content or fatty acid composition in *Brassica napus*. *Plant J.* 2020;104:1410–22
46. Zhang T, Dong X, Yuan X. *et al.* Identification and characterization of CsSRP43, a major gene controlling leaf yellowing in cucumber. *Hortic Res.* 2022;9:uhac212
47. Pendleton M, Sebra R, Pang AWC. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods.* 2015;12:780–6
48. Grob S, Schmid MW, Grossniklaus U. Hi-C Analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the flamenco locus of *Drosophila*. *Mol Cell.* 2014;55:678–93
49. Cheng H, Concepcion GT, Feng X. *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods.* 2021;18:170–5
50. Durand NC, Shamim MS, Machol I. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 2016;3:95–8
51. Dudchenko O, Batra SS, Omer AD. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* 2017;356:92–5
52. Robinson JT, Turner D, Durand NC. *et al.* Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* 2018;6:256–258.e1

53. Manni M, Berkeley MR, Seppey M. *et al.* BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38:4647–54
54. Manni M, Berkeley MR, Seppey M. *et al.* BUSCO: assessing genomic data quality and beyond. *Curr Protoc.* 2021;1:e323
55. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics.* 2003;00:10.3.1–10.3.18
56. Marçais G, Delcher AL, Phillippy AM. *et al.* MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol.* 2018;14:e1005944
57. Goel M, Sun H, Jiao W-B. *et al.* SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 2019;20:277
58. Flynn JM, Hubley R, Goubert C. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 2020;117:9451–7
59. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* 2009;25:4.10.1–4.10.14
60. Li Z, Zhang Z, Yan P. *et al.* RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics.* 2011;12:540
61. Grabherr MG, Haas BJ, Yassour M. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol.* 2011;29:644–52
62. Haas BJ, Delcher AL, Mount SM. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 2003;31:5654–66
63. Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform.* 2020;2:lqaa026
64. Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol Biol.* 2019;1962:161–77
65. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics.* 2011;12:491
66. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods.* 2021;18:366–8
67. Jones P, Binns D, Chang HY. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30:1236–40
68. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol.* 2016;428:726–31
69. Calle García J, Guadagno A, Paytuvi-Gallart A. *et al.* PRGdb 4.0: an updated database dedicated to genes involved in plant disease resistance process. *Nucleic Acids Res.* 2021;50:D1483–90
70. Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol.* 1985;2:150–74
71. Retief JD. Phylogenetic analysis using PHYLIP. In: Misener S, Krawetz SA, eds. *Bioinformatics Methods and Protocols.* Humana Press: New Jersey, 1999,243–58
72. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–100
73. Hao Z, Lv D, Ge Y. *et al.* Rldeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput Sci.* 2020;6:e251
74. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 2007;35:W265–8
75. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics.* 2008;9:18
76. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 2018;176:1410–22
77. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999;27:573–80
78. Lin Y, Ye C, Li X. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res.* 2023;10:uhad127
79. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754–60
80. Li H, Handsaker B, Wysoker A. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25:2078–9
81. Genovese G, Rockweiler NB, Gorman BR. *et al.* BCftools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics.* 2024;40:btac038
82. Rausch T, Zichner T, Schlattl A. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28:i333–9
83. Lichtenthaler HK. Chlorophylls and carotenoids: pigments of photosynthetic biomembranes. In: Packer L (ed.), *Methods in Enzymology*, Vol. 148. Plant Cell Membranes. San Diego: Academic Press, 1987,350–82
84. Sun W, Li X, Huang H. *et al.* Mutation of CsARC6 affects fruit color and increases fruit nutrition in cucumber. *Theor Appl Genet.* 2023;136:111
85. Lichtenthaler HK, Buschmann C. Chlorophylls and carotenoids: measurement and characterization by UV-VIS spectroscopy. In: Wrolstad RE, *et al.* (ed.), *Current Protocols in Food Analytical Chemistry*. New York: John Wiley & Sons, 2001,F4.3.1–8
86. McKenna A, Hanna M, Banks E. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303
87. Hill JT, Demarest BL, Bisgrove BW. *et al.* MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. *Genome Res.* 2013;23:687–97
88. Xi X, Wei K, Gao B. *et al.* BrFLC5: a weak regulator of flowering time in Brassica rapa. *Theor Appl Genet.* 2018;131:2107–16
89. Fang L, Wei XY, Liu LZ. *et al.* A tobacco ringspot virus-based vector system for gene and microRNA function studies in cucurbits. *Plant Physiol.* 2021;186:853–64
90. Xu S, Wang Y, Yang S. *et al.* The WUSCHEL-related homeobox transcription factor CsWOX3 negatively regulates fruit spine morphogenesis in cucumber (*Cucumis sativus* L.). *Hortic Res.* 2024;11:uhae163