nature genetics

Article

https://doi.org/10.1038/s41588-025-02126-0

The genome and GeneBank genomics of allotetraploid *Nicotiana tabacum* provide insights into genome evolution and complex trait regulation

Received: 6 June 2023

Accepted: 13 February 2025

Published online: 26 March 2025

Check for updates

Yanjun Zan[1,11] ✉, Shuai Chen[1,11], Min Ren[1,11], Guoxiang Liu[1,11], Yutong Liu[1,11], Yu Han[2], Yang Dong[3,4], Yao Zhang[3,5], Huan Si[1], Zhengwen Liu[1], Dan Liu[1], Xingwei Zhang[1], Ying Tong[1], Yuan Li[1], Caihong Jiang[1], Liuying Wen[1], Zhiliang Xiao[1], Yangyang Sun[1], Ruimei Geng[1], Yan Ji[1,6], Quanfu Feng[1], Yuanying Wang[1], Guoyou Ye[7,8], Lingzhao Fang[9], Yong Chen[10] ✉, Lirui Cheng[1] ✉ & Aiguo Yang[1] ✉

Nicotiana tabacum is an allotetraploid hybrid of *Nicotiana sylvestris* and *Nicotiana tomentosiformis* and a model organism in genetics. However, features of subgenome evolution, expression coordination, genetic diversity and complex traits regulation of *N. tabacum* remain unresolved. Here we present chromosome-scale assemblies for all three species, and genotype and phenotypic data for 5,196 *N. tabacum* germplasms. Chromosome rearrangements and epigenetic modifications are associated with genome evolution and expression coordination following polyploidization. Two subgenomes and genes biased toward one subgenome contributed unevenly to complex trait variation. Using 178 marker–trait associations, a reference genotype-to-phenotype map was built for 39 morphological, developmental and disease resistance traits, and a novel gene regulating leaf width was validated. Signatures of positive and polygenic selection during the process of selective breeding were detected. Our study provides insights into genome evolution, complex traits regulation in allotetraploid *N. tabacum* and the use of GeneBank-scale resources for advancing genetic and genomic research.

Nicotiana tabacum (common tobacco, $2n = 4x = 48$) is an interspecific hybrid of two progenitors, *Nicotiana sylvestris* ($2n = 2x = 24$) and *Nicotiana tomentosiformis* ($2n = 2x = 24$), that merged approximately 0.2 million years ago[1–3]. Such whole-genome duplications are typical of all land plants and are thought to have made adaptive contributions in times of global catastrophe. Although the most common fate of polyploids appears to be fractionation and eventual reversion to the diploid state, our understanding of genome evolution following whole-genome duplication remains incomplete. Various mechanisms, such as homoeologous chromosome exchange[4–6], reactivation of transposable elements (TEs)[7] and DNA methylation repatterning[8,9], have been reported, yet findings from different studies are often controversial[4–9]. In the case of *N. tabacum*, obtaining a complete assembly for *N. tabacum* and its progenitor constitutes a substantial challenge because of difficulties in resolving highly repetitive repeat regions and in computationally disentangling homoeologous sequences with high similarity from

A full list of affiliations appears at the end of the paper. ✉e-mail: zanyanjun@caas.cn; chycht@163.com; chenglirui@caas.cn; yangaiguo@caas.cn

two progenitor species[1–3]. As a result, the impact of these previously reported processes on the genome and transcriptome of *N. tabacum* remains poorly understood. Nearly all commercial tobaccos belong to the species *N. tabacum*, with more than 7,000 varieties cataloged in the Plant Germplasm system of the United States Department of Agriculture and China. They are commonly classified as flue-cured (for cigarettes), sun-cured (for pipe smoking), burley (blends for pipe smoking), cigar (for cigars) or oriental (blends) based on agricultural practices[10]. They have distinctive characteristics in terms of plant architecture, leaf morphology and metabolic traits, which differ significantly from those of their ancestors. Previous studies have revealed several quantitative trait loci (QTLs) associated with nicotine content, leaf shape, plant height and disease resistance traits[11–16]. However, the genetic differentiation among major types of *N. tabacum* and how variations from two subgenomes contribute to the remarkable phenotypic diversity are unclear. A deeper understanding of these questions could provide insights into the evolution and adaptation of polyploids, as well as the genetic regulation of economically important traits in polyploid species, such as wheat, cotton and *Brassica napus*.

Recent advances in sequencing have enabled cost-effective assembly and GeneBank-scale sequencing of crops with large and complex genomes[17,18]. Here we present a complete chromosome-level assembly of the allotetraploid *N. tabacum* genome along with its ancestral genomes, genotype and phenotype data for an entire *N. tabacum* GeneBank with 5,196 germplasms. Through comparative genomic, transcriptomic and epigenomic analysis as well as genome-wide association analyses, we revealed biased genome downsizing toward the T subgenome, epigenetic modifications associated with subgenome expression divergence and an uneven contribution of the two subgenomes to complex trait variation. Our findings, together with released data and seed stocks, will likely accelerate aspects of tobacco research that were previously hindered by the complexity of the polyploid genome, with benefits in plant genomic and genetic research.

Results

Genome evolution highlighted by chromosome-scale assemblies

Based on *k*-mer analysis with Illumina short reads and flow cytometry analysis, the genomes of *N. tabacum*, *N. sylvestris* and *N. tomentosiformis*[3] were estimated to be approximately 4.38, 2.38 and 2.24 Gb, respectively (Fig. 1a and Supplementary Figs. 1 and 2). To overcome the challenge of assembling a large polyploid genome, we generated 52X Illumina short reads, 123.35X PacBio Sequel II reads, 124.93X 10X Genomic linked reads and 120.12X Hi-C reads from *N. tabacum* L. var. ZY300 and adopted a hybrid assembly approach (Methods). The final assembly included 4.17 Gb (Fig. 1b) of sequences with a contig N50 of 27.17 Mb, and 96.98% of the sequences were anchored to 24 pseudo-chromosomes (Supplementary Fig. 1a,d and Supplementary Table 1). An almost complete telomere-to-telomere assembly of the *N. sylvestris* genome with three gaps was obtained with 36X PacBio Revio reads, 180.59X Hi-C reads and 42X Illumina short reads. The final assembly contained a 2.38-Gb sequence with a contig N50 of 190 Mb (Fig. 1c, Supplementary Fig. 1b,e and Supplementary Table 1). By contrast, assembly of the *N. tomentosiformis* genome was complicated by repetitive regions. With 40X PacBio Revio reads, 80X ONT reads, 138X Hi-C reads and 50X Illumina short reads, our assembly still included five complex regions, amounting to approximately 500 Mb, with abnormal Hi-C signals (Supplementary Fig. 1f). Detailed investigations revealed that these regions were characterized by a high level of repeats and a high density of TEs (Fig. 1d). Because these regions were assembled with ONT and HiFi reads by the Hifiasm assembler and genome coverage was even after Illumina and HiFi reads were mapped to the assembled genome (Supplementary Fig. 3), we retained these regions in the final assembly of 2.24 Gb of sequences with a contig N50 of 170.53 Mb (Fig. 1d) and five gaps. Detailed evaluations and comparisons showed that these three genome assemblies represented 10.51%, 6.59% and 24.75% increases in the assembled genome size and 77.02-fold, 2,383.67-fold and 2,064.62-fold increases in the contig N50 size, respectively, compared with those of the latest assemblies (Supplementary Table 2 and Supplementary Note)[1,3,16].

A total of 80,433, 40,290 and 37,862 protein-coding genes were annotated in the *N. tabacum*, *N. sylvestris* and *N. tomentosiformis* genomes, respectively (Supplementary Table 3), of which 95.36%, 98.82% and 99.23%, respectively, could be functionally annotated (Supplementary Table 4 and Supplementary Note). A total of 3,349,090,085, 1,782,416,235 and 1,823,429,503-bp repeat elements were identified, accounting for 82.75%, 74.92% and 81.26% of the assembled *N. tabacum*, *N. sylvestris* and *N. tomentosiformis* genomes, respectively (Supplementary Table 5). As mentioned above, the T subgenome contains more repetitive sequences than the S subgenome, and these sequences are enriched on chromosome (chr.) 2, chr. 17 and chr. 21 (Supplementary Figs. 1 and 4). Further analysis revealed a significantly greater proportion of retrotransposons than DNA transposons in these regions (Supplementary Fig. 4). Similar to other plants, long terminal repeat retrotransposons (LTRs) (66.21–70.18%) were the predominant retrotransposons, with Gypsy (41.78–43.40%) being the most abundant (Supplementary Table 5) for these genomes. For *N. tabacum*, 14 centromeres were pinpointed by combining signals from CENH3 chromatin immunoprecipitation sequencing (ChIP-seq) in our study and previous reports[19–21] as well as de novo centromere prediction based on tandem repeat monomers using quarTeT[22]. Except for chr. 16, chr. 18 and chr. 22, where evidence from the three analyses was controversial, candidate centromeric regions were identified for all the remaining chromosomes (Fig. 1e, Supplementary Fig. 5 and Supplementary Table 6).

The genome of *N. tabacum* was in large blocks of synteny to the genomes of the ancestral species (Fig. 1f and Supplementary Fig. 6). The majority of the blocks were collinear, except for 1,420 inversions, 539 duplications and 725 deletions (size >1 Mb) (Supplementary Fig. 6 and Supplementary Table 7). In total, 56.99% and 43.01% of the genome was partitioned to *N. sylvestris* (S) and *N. tomentosiformis* (T), respectively, with 11 chromosome rearrangement events pinpointed at approximately 1-kb resolution (Supplementary Table 8). For example, homoeologous chromosome exchange between *N. sylvestris* chr. 18 and *N. tomentosiformis* chr. 9 generated chr. 9 and chr. 18 for *N. tabacum* (Fig. 1f,g). This suggested that chromosomal rearrangements produced changes in genome structure, which likely stabilized chromosome pairing during meiosis. Further comparative genomic analysis between our assemblies and previously reported assemblies revealed that all of these homoeologous exchange events were conserved, suggesting that they likely resulted from polyploidization rather than intraspecies variation (Supplementary Note). Based on a genome survey of multiple individuals, the genome sizes of *N. sylvestris, N. tomentosiformis and N. tabacum* were 2.32 ± 0.07, 2.14 ± 0.04 and 4.31 ± 0.07 Gb, respectively. Given this, the genome of the allotetraploid *N. tabacum* likely decreased by approximately 3.45%, which is consistent with the previously published decrease of 3.7% (refs. 1,3). Nearly all segments (98%) present in the *N. sylvestris* genome were found in the *N. tabacum* genome, while the repetitive regions located on chr. 2, chr. 17 and chr. 21 of the *N. tomentosiformis* genome were absent (Fig. 1f and Supplementary Fig. 6). Together with previously reported biased downsizing toward repetitive sequences from the T genome using 454 sequencing data[23], these results revealed that genetic changes occur more rapidly in the T subgenome than in the S subgenome and that genome downsizing is strongly biased toward the T genome at these repetitive regions. Estimation of the LTR insertion time via the Kimura distance method[24] revealed that approximately 7.56% of the LTRs were inserted after the two subgenomes merged 0.2 million years ago (Fig. 1h). Although dating the exact times of polyploidization and LTR insertion make heavy assumptions that are challenging to evaluate, this result suggests that polyploidization likely did not stimulate extensive reactivation of LTRs.

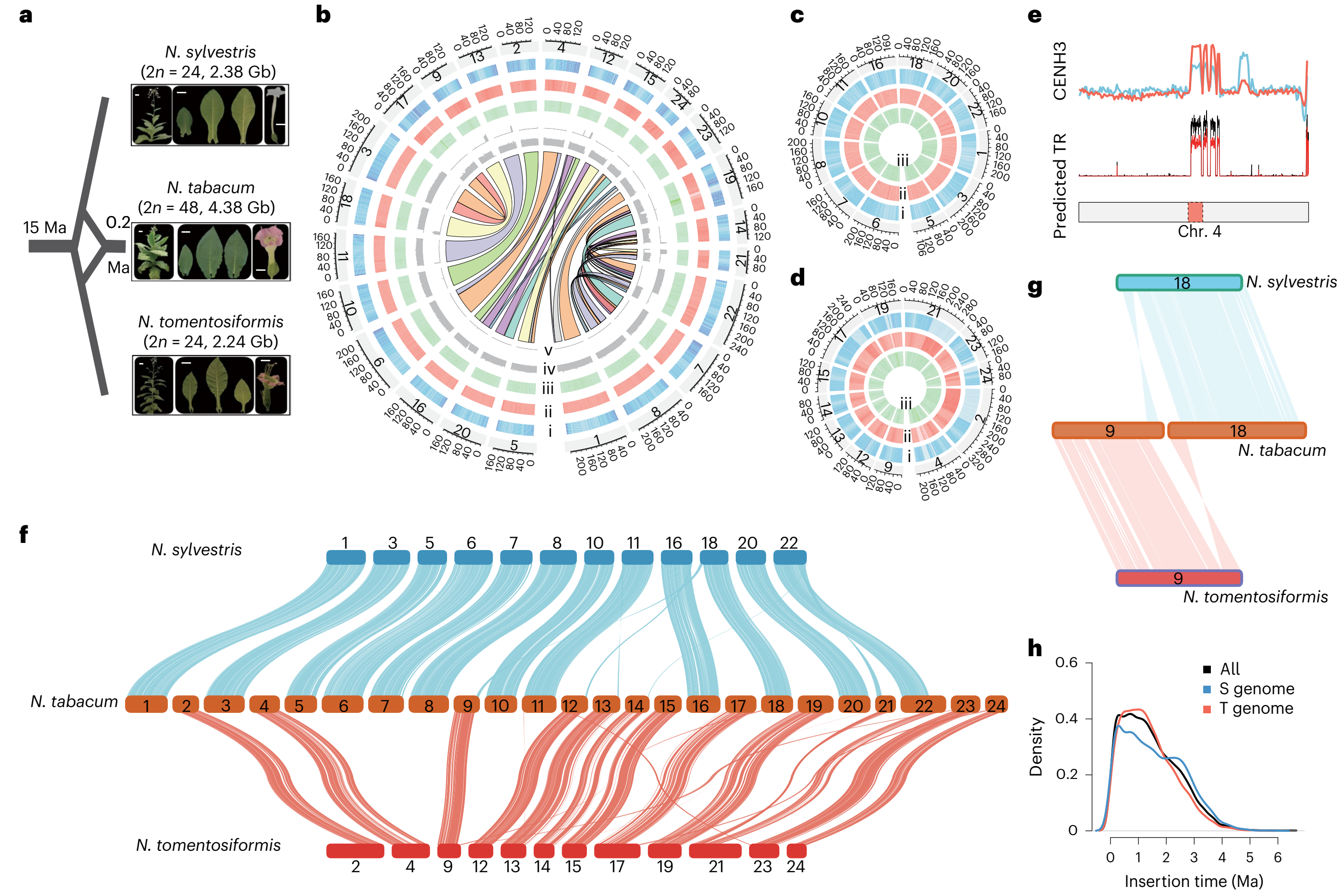

Fig. 1 | High-quality assembly and comparative genomic analysis of the *N. tabacum, N. sylvestris and N. tomentosiformis* genomes. a, Schematic illustration of the origin of *N. tabacum* and phenotype diversity among two diploid progenitors, *N. sylvestris* and *N. tomentosiformis*, and one of the allotetraploid *N. tabacum* samples. Scale bars, 20 cm (whole plant), 20 cm (leaf) and 1 cm (flower). **b**–**d**, Genomic features of *N. tabacum* (**b**), *N. sylvestris* (**c**) and *N. tomentosiformis* (**d**). The following tracks from the outermost and innermost regions were identified: gene density (i), TE density (ii), GC content (iii), and sequence coverage obtained by mapping Illumina short reads from *N. sylvestris* (iv) and *N. tomentosiformis* (v) to *N. tabacum*. Synteny blocks between the two subgenomes are represented by colored linkers across the center of the plot. **e**, Locations of the centromeric regions on chr. 4. The red and blue lines indicate two replicates of CENH3 ChIP-seq signals. TR indicates the density of typical centromeric repeats obtained from CENH3 ChIP-seq in previous studies. The predicted centromeric regions are based on tandem repeat monomers. **f**, Genome alignment between *N. tabacum* and *N. sylvestris* as well as between *N. tabacum* L. var. ZY300 and *N. tomentosiformis*. **g**, Example of a homoeologous chromosome exchange event. **h**, Estimated insertion time for an LTR. Ma, million years ago.

Subgenome expression divergence and epigenetic modification

To determine gene expression and DNA methylation evolution in *N. tabacum*, we generated RNA sequencing (RNA-seq) and bisulfate sequencing (bisulfate-seq) data for the two ancestors and *N. tabacum* (Methods, Supplementary Note and Supplementary Table 9). The overall expression of *N. tabacum* genes was similar to that of their ancestors (mean fold change = 0.01 for the S versus S subgenome and 0.04 for the T versus T subgenome) (Supplementary Fig. 7), with only 2,025 and 2,040 genes differentially expressed between the S and S subgenomes and between the T and T subgenomes, respectively ($\log_2$(fold change) > 2, false discovery rate < 0.05) (Fig. 2a,b and Supplementary Table 10). Notably, when gene expression was upregulated in the allotetraploid *N. tabacum*, we found significantly decreased CHG methylation levels and vice versa, suggesting that changes in gene expression following polyploidization are associated with epigenetic modifications. For example, among the 2,025 genes differentially expressed between the S and S subgenomes (Fig. 2a), 1,302 were upregulated in the S subgenome (red dots in Fig. 2a). Both CG and CHG methylation levels were lower in the S subgenome (Fig. 2c), indicating that polyploidization likely increased gene expression by reducing methylation levels. A similar epigenetic modification pattern was observed for genes downregulated in the S subgenome (Fig. 2d), upregulated in the T subgenome (Fig. 2e) and downregulated in the T subgenome (Fig. 2f). Next, we compared the expression of 28,143 unique homoeologous gene pairs (originating from speciation and brought back in the same genome by allopolyploidization) between the two subgenomes and homologous gene pairs (derived from different parental species but related by ancestry) between the two ancestors to evaluate homoeologous and homologous gene expression patterns. Although no evidence for overall homoeologous gene expression bias (HEB) was found (median $\log_2$(fold change) = 0.05 for *N. sylvestris* versus *N. tomentosiformis* homologous gene pairs and 0.01 for the S versus T subgenome homoeologous gene pairs) (Supplementary Fig. 7), 5,753 homoeologous gene pairs were differentially expressed between the two *N. tabacum* subgenomes ($\log_2$(fold change) > 2, false discovery rate < 0.05) (Supplementary Table 11). The majority (4,073, 70.70%) were inherited from ancestors as a parental legacy, while the remaining 1,680 (29.30%) pairs were likely triggered by polyploidization (Fig. 2g). In addition, 2,284 gene pairs were differentially expressed between the two ancestors but not between the two subgenomes of *N. tabacum* (Fig. 2g (set 4) and Fig. 2h (set 8)), indicating that polyploidization not

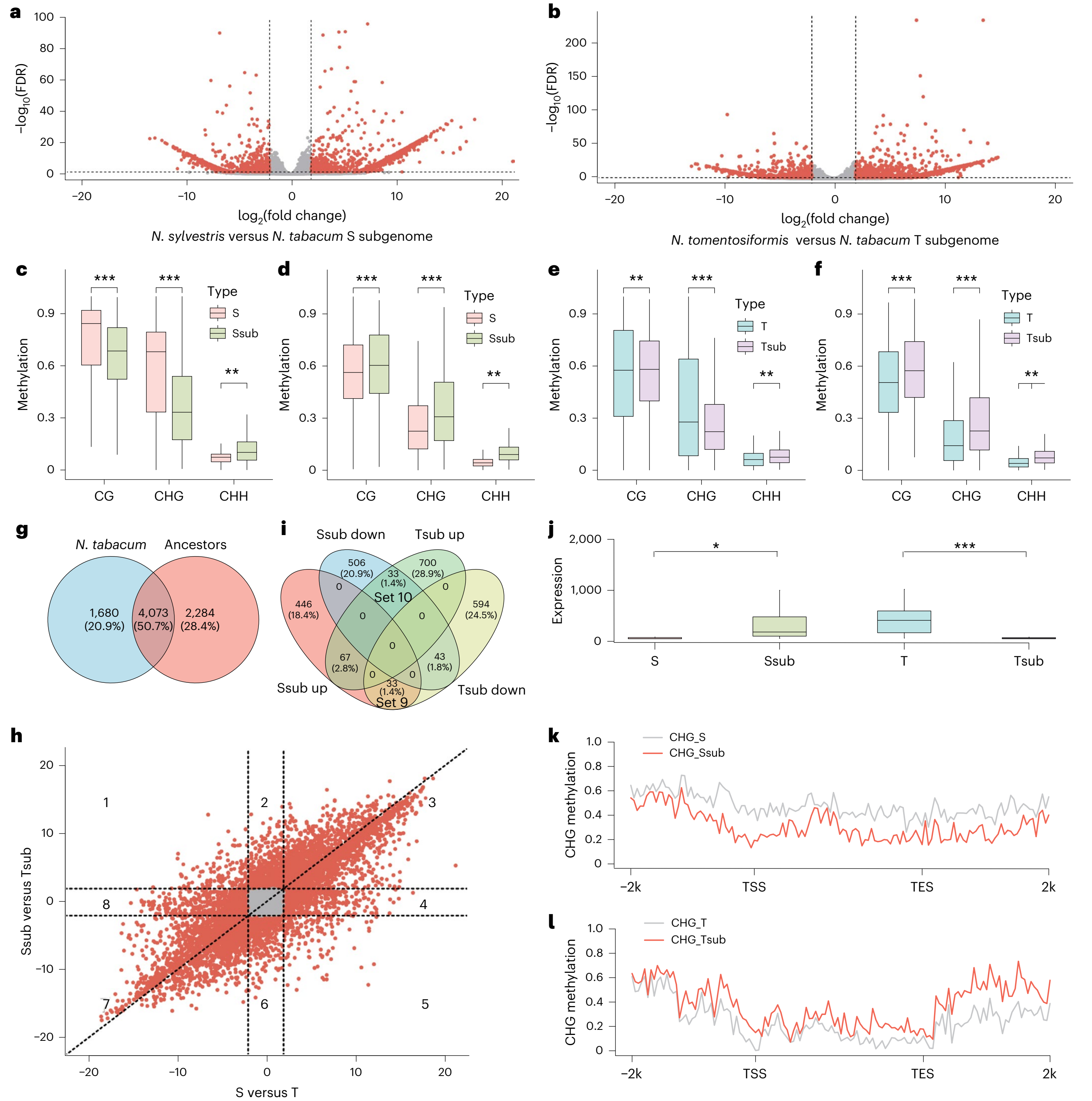

Fig. 2 | Subgenome gene expression divergence is associated with epigenetic modifications. a,**b**, Gene expression divergence between the *N. sylvestris* and *N. tabacum* S subgenomes (**a**) and between the *N. tomentosiformis* and *N. tabacum* T subgenomes (**b**). **c**–**f**, CG, CHG and CHH methylation differences among ancestral genomes and the corresponding *N. tabacum* subgenome for genes upregulated (**c**) or downregulated (**d**) in the S subgenome and upregulated (**e**) or downregulated (**f**) in the T subgenome. Sample size is 1,302 (**c**), 723 (**d**), 1,236 (**e**) and 804 (**f**). *P* values from left to right are 1.98×10^{-17}, 2.20×10^{-72}, 6.81×10^{-3}, 2.28×10^{-4}, 2.44×10^{-14}, 7.06×10^{-3}, 9.77×10^{-3}, 1.00×10^{-15}, 2.04×10^{-3}, 2.90×10^{-11}, 1.80×10^{-19} and 6.23×10^{-3}. **g**, Overlap of differentially expressed homoeologous gene pairs between the two subgenomes of *N. tabacum* (blue) and two ancestor genomes (red). **h**, Relationships of homologous or homoeologous gene expression fold changes between two subgenomes of *N. tabacum* (*y* axis) and two ancestor genomes (*x* axis). Dashed vertical and horizontal lines indicate $\log_2$(fold change) > 2. **i**, Overlap of homoeologous gene expression changes. The overlap between Ssub up and Tsub up indicates that a pair of homoeologous genes were both upregulated in the *N. tabacum* subgenome, while the overlap between Ssub down and Tsub up indicates that one gene from the S subgenome was downregulated but its homolog was upregulated in the T subgenome. **j**, Boxplot illustrating that homoeologous gene expression is strongly biased toward the S subgenome by simultaneously upregulating genes in the S subgenome and suppressing their homoeologs in the T subgenome. The sample size is 33 and *P* values are 2.21×10^{-2} and 1.52×10^{-5} from left to right. **k**,**l**, Comparison of CHG methylation levels between the ancestor genome and corresponding subgenomes for homoeologous genes strongly biased toward the S (**k**) and T (**l**) genomes. In the boxplots, the center line is the median, box limits are the first and third quartiles and whiskers are the minimum and maximum. A two-sided *t*-test was performed to generate the significances for all pairwise comparisons. *$P<0.05$, **$P<0.01$ and ***$P<0.001$. 2k, 2 kilobase; FDR, false discovery rate; TES, transcript end site; TSS, transcript starting site.

only leads to gene expression divergence but also equalizes differences. Notably, there were 66 homoeologous gene pairs (Fig. 2i (sets 9 and 10) and Supplementary Table 12) for which one of the homoeologous genes was upregulated or downregulated, and the counterpart from the other subgenome was changed to the opposite direction (Fig. 2j–l). For example, 33 genes were significantly upregulated between the S and S subgenomes ($P = 2.21 \times 10^{-2}$, $\log_2$(fold change) > 2; Fig. 2i,j (set 9)), while their homoeologous genes were significantly downregulated between the T and T subgenomes ($P = 1.52 \times 10^{-5}$). Meanwhile, the CHG methylation level significantly decreased ($P = 2.91 \times 10^{-24}$) in the S subgenome and increased in the T subgenome ($P = 2.56 \times 10^{-10}$) (Fig. 2j–l). This indicates that gene expression for these homoeologous pairs was biased toward one subgenome by simultaneously upregulating expression in one subgenome and suppressing another subgenome, which was likely mediated by epigenetic modifications at contrasting directions. Gene Ontology enrichment analysis revealed that homoeologous genes biased toward the S subgenome and T subgenome were enriched in defense response, flowering development, cell communication, signaling transduction, and so on (Supplementary Fig. 8).

Genetic diversity of a global collection of 5,196 lines

To understand the global distribution of genetic variation and differentiation among major tobacco types (Fig. 3a,b, Supplementary Fig. 9 and Supplementary Table 13), we genotyped an entire GeneBank collection hosted at the Chinese Academy of Agriculture Sciences using a genotype-by-sequencing approach (Methods, Supplementary Table 14 and Supplementary Note). This includes 2,582 sun-cured tobacco, 2,152 flue-cured tobacco, 223 burley tobacco, 126 cigar tobacco and 113 oriental tobacco germplasms, giving to 5,196 germplasms in total.

Using 95,308 single-nucleotide polymorphisms (SNPs) covering 98% of the genomic bins at 1-Mb resolution (Fig. 3c), we found that geographic origin at the continental scale was the most important correlate of genetic structure and that genetic structure did not always match the conventional types defined on the basis of agronomical practices. For example, the first principal component separated flue-cured tobacco from sun-cured tobacco to a large extent. However, the majority of samples located on the left-hand side of Fig. 3d were of flue-cured tobacco originating from North America or derived materials belonging to the North American lineage, while plants located on the opposite side were sun-cured tobacco landraces collected in China (Fig. 3d,e). In addition, a latitudinal cline was observed along the second principal component. Sun-cured tobacco landraces from northern China with a substantially shorter life cycle, colder growth temperature and longer day length were located in the upper right corner (Fig. 3d), while plants sampled in southern China with opposite climate signatures clustered at the bottom of Fig. 3d. According to historical records, burley tobacco originated from a single mutated variety, known as White Burley, characterized by a chlorophyll deficiency phenotype controlled by double homozygous recessive alleles at the Yellow burley 1 (*YB1*) and Yellow burley 2 (*YB2*) loci[25]. Consistent with this, burley tobacco samples were clustered together near the North American group and showed the highest pairwise identity-by-state (IBS) value (median = 0.88) (Fig. 3f). Cigar and oriental tobacco samples were scattered among sun-cured and flue-cured tobacco samples, indicating a likely history of mixed breeding (Fig. 3d,e).

Pairwise IBS values for the five tobacco types ranged from 0.75 to 0.88, suggesting that the gene pool used to develop each type of tobacco was very narrow. However, the distribution of IBS values for flue-cured, sun-cured and cigar tobacco was multimodal, indicating that the co-occurrence of divergent gene pools was likely caused by introgression from imported materials. Despite the marked phenotypic differentiation in plant morphology, flowering and metabolic traits (Fig. 3b and Supplementary Fig. 9), genome-wide differentiation among the five types and three genetic groups was low, with pairwise population differentiation coefficient (F_{ST}) values ranging from 0.04 to 0.23 (Fig. 3g,h) and no selective sweeps were detected. This was likely due to the short and special selective breeding history of less than a few hundred years, which aimed to balance quality among many leaf characteristics.

Subgenome divergence and regulation of complex traits

With the availability of genome-wide markers and phenotypic measurements for more than 5,000 lines, we applied the genome-wide association study (GWAS) approach to generate a comprehensive catalog of the allelic variation underlying differences in 43 traits (Figs. 3a and 4a) with intermediate to high heritability ($h^2 = 0.15$–0.82) (Supplementary Tables 15 and 16). A total of 178 significant marker–trait associations were identified for 39 traits (Fig. 4b and Supplementary Fig. 10), and several high-potential associations were detected (Fig. 5 and Supplementary Note). *P* values and additive effects for the QTLs are summarized in Supplementary Table 17. Although linkage disequilibrium (LD) is too extensive to directly pinpoint the candidate genes underlying each association peak (LD decay to 0.2 within 500 kb, givens eight genes on average) (Supplementary Fig. 11), we found a few genes previously identified as QTLs[11–16] or annotated as potential candidates near eight peaks (dashed lines in Fig. 4b). In addition, we conducted a literature review[11–16] to determine the relative position of previously reported QTLs (Supplementary Table 18) in the ZY300 genome assembly and discovered that 95% of the QTLs detected in our study were novel. Taken together, these marker–trait associations constitute a comprehensive genotype-to-phenotype map (Fig. 4d and Supplementary Table 19), providing a roadmap for future genetic studies in this species.

For several traits, such as flowering time (FT) and resistance to black shank (RBSH), we detected QTLs from two homoeologous chromosomes, whereas only one QTL from one subgenome was detected for the remaining traits (Fig. 4 and Supplementary Note). Therefore, we explored subgenome divergence in the regulation of 28 continuously distributed complex traits (Fig. 4e) by partitioning phenotypic variance into two subgenomes by fitting a joint mixed model with two random effects, each estimated from markers in the corresponding subgenome. Except for RBSH, the T subgenome explained disproportionately more of the variation in all the remaining disease resistance traits (median ratio of variance explained by Ssub/Tsub = 0.55) (Fig. 4e). Moreover, the S genome made a greater contribution to budding time, RBSH and leaf number, with the ratios of variance explained by the Ssub/Tsub genome being 1.55, 1.48 and 1.37 (Fig. 4e), respectively. To explore the role of 3,964 homoeologous gene pairs (Fig. 2g), whose HEB was either triggered or equalized by polyploidization, in complex trait variation, we partitioned phenotypic variance into genomic regions harboring these genes and the remaining part of the genome (non-HEB). Although there were only 2,649 (2.78%) markers near the 5-kb regions of these genes, they made disproportionately larger contributions to the variation in resistance to powdery mildew (ratio = 6.08), FT (ratio = 1.52) and so on (Fig. 4f). These analyses highlight the unequal contributions of the two subgenomes to complex trait variation and provide potential evidence for the role of HEB in the variation of complex traits.

Arf9 is associated with leaf width variation

Strong agronomical interest in leaf usage resulted in an excess of diversity in leaf morphology (Fig. 4a). Here we detected 65 QTLs for 12 leaf morphology traits (Supplementary Table 17), including leaf width (LW), leaf length, vein diameter, leaf stem angle, leaf vein angle, leaf auricle, leaf thickness, leaf shape, leaf tip, leaf color, leaf serration and leaf flatness (LF). Overall, the correlations among these traits were relatively low (median Spearman correlation = 0.05) (Supplementary Fig. 12). However, three QTLs were simultaneously associated with LW and LF (Figs. 4b and 5a). One of these QTLs (chr. 23:148211202) is located at the end of chr. 23. Allele T at this locus increased the LW by 1.11 ± 0.18 ($P = 1.66 \times 10^{-9}$) and increased the LF by 0.2 ± 0.02 units ($P = 1.58 \times 10^{-18}$) (Supplementary Fig. 13). Like the peak described in

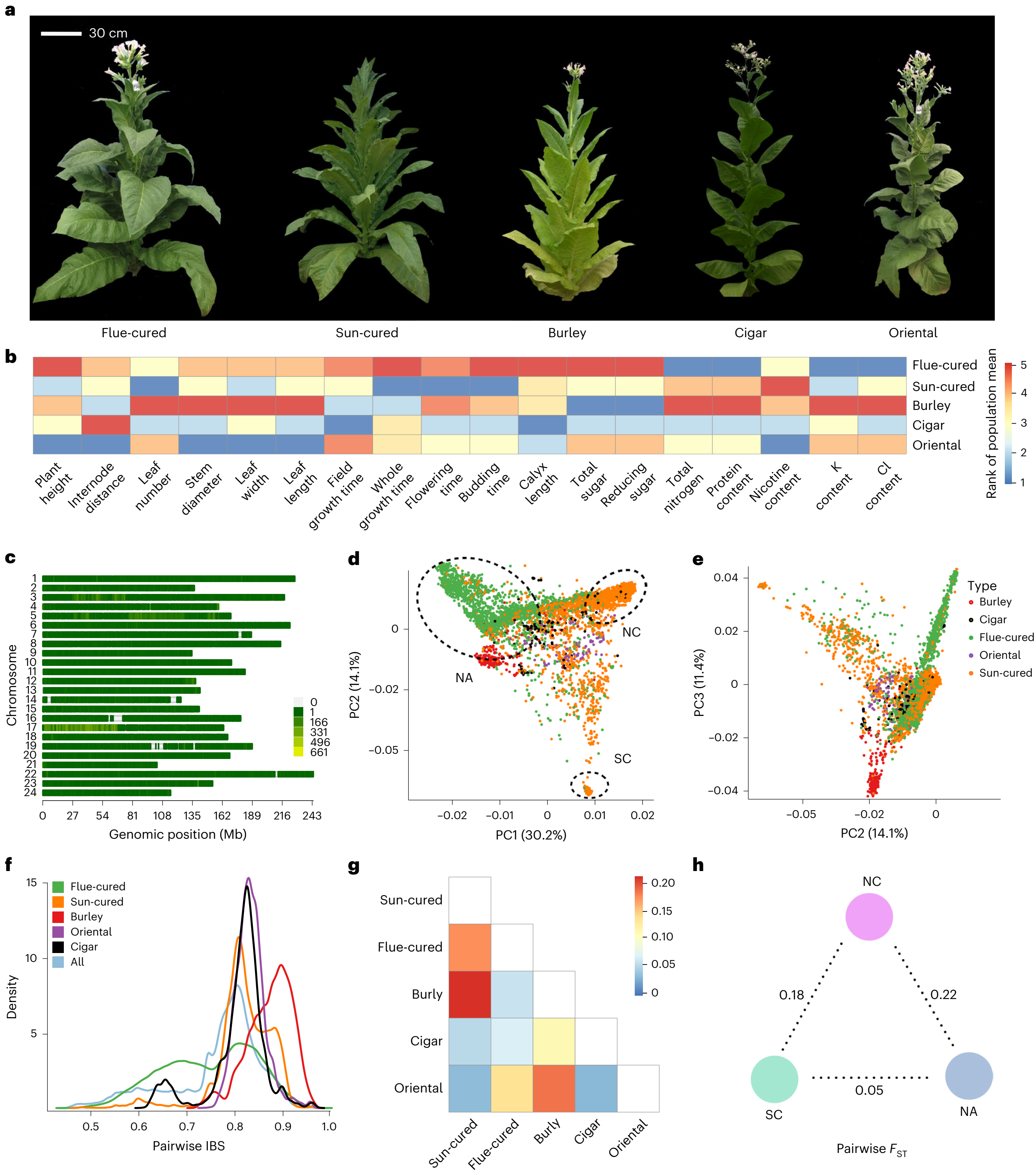

Fig. 3 | Genetic and phenotypic diversity of a global collection of 5,196 tobacco germplasm resources. a, Five representative individuals of five types of tobacco plants. **b**, Heatmap of the mean phenotypic rank for 20 traits (*x* axis) among five types of tobacco named after their role in agronomical practices. Mean phenotype ranks were obtained by first calculating the phenotype mean in each group and then ranking among the five groups (by column). **c**, Genomic distribution of 95,308 SNPs in the *N. tabacum* genome identified using the genotype-by-sequencing approach. *x* axis is the genomic position and *y* axis is the corresponding chromosome. **d**,**e**, Genetic structure of 5,196 *N. tabacum* lines inferred from principal component analysis with the first and second principal component illustrated in **d**, while the second and third component illustrated in **e**. The samples are colored according to the type assigned on the basis of agronomic practices. **f**, Distribution of pairwise IBS values in different tobacco types. **g**,**h**, F_{ST} between different tobacco types (**g**) and groups defined on the basis of genetic structure (**h**). PC1, first principal component; PC2, second principal component. NA, North America; NC, North China; SC, South China.

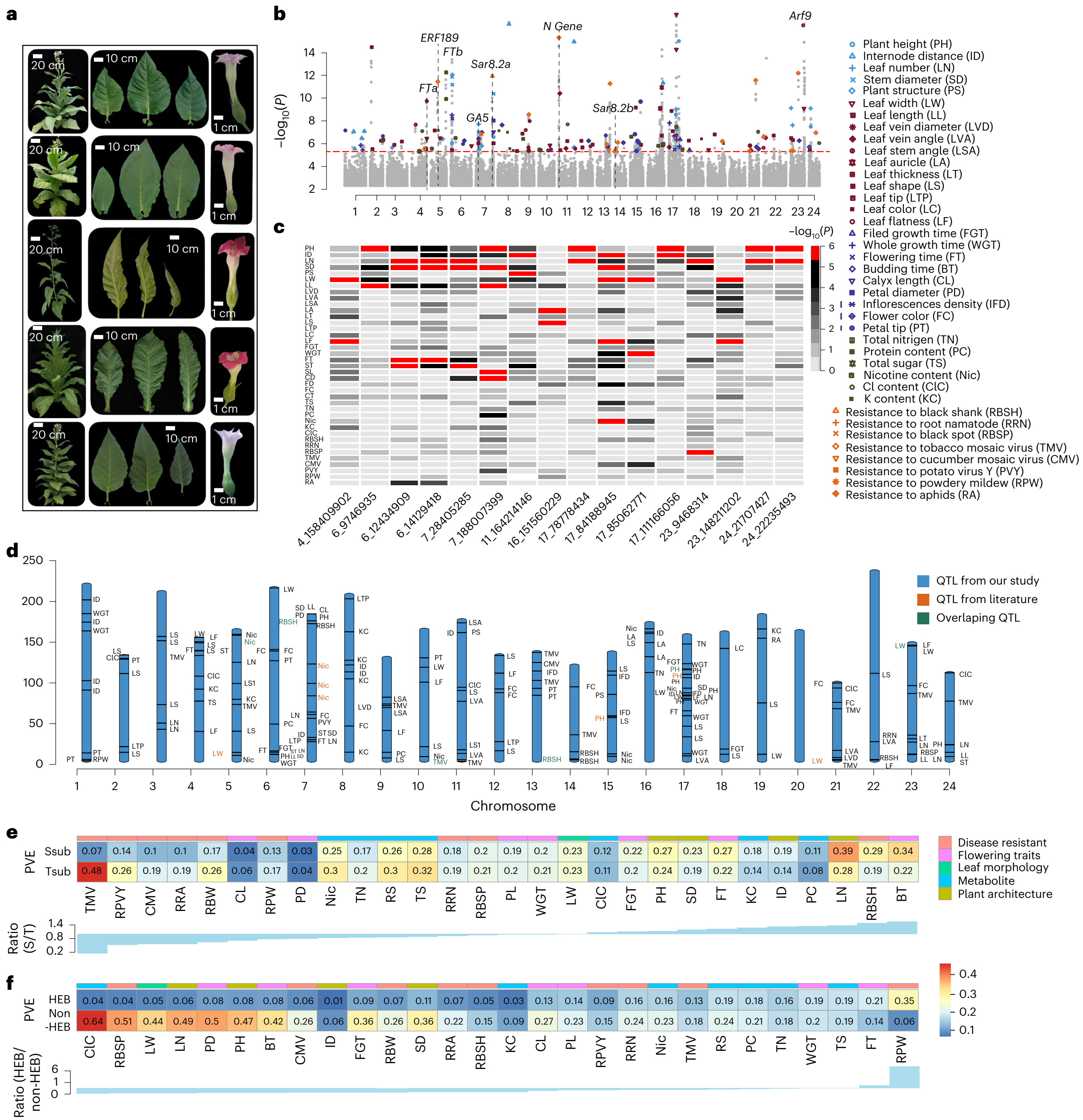

Fig. 4 | Association results for 43 plant morphological, physiological, metabolic and disease resistance traits, and the contributions of the subgenome and HEB to complex trait variation. **a**, Phenotypic diversity of plant architecture, leaf morphology and flowering morphology and color among five selected plants. The vertical dashed gray lines highlight the genomic positions of potential candidate genes. **b**, Manhattan plots of results from 43 GWAS scans. The red horizontal dashed lines indicate the Bonferroni-corrected genome-wide significance thresholds. The vertical dashed gray lines highlight the genomic positions of 16 SNPs associated with more than two traits. **c**, Heatmap illustrating the *P* values of 16 SNPs detected for more than two traits. Each cell represents $-\log_{10}(P)$ of a particular SNP (*x* axis) associated with a specific trait (*y* axis on the right). **d**, The reference genotype-to-phenotype map includes marker–trait associations detected in our study (orange), in previous reports (blue) and in both (green). Previously reported QTLs spanning more than 10 Mb and QTLs without probe sequences are not included here. **e**, Unequal contribution from two subgenomes to complex trait variation. **f**, The contribution of homoeologous genes whose expression bias (HEB) was either triggered or equalized by polyploidization to complex trait variation.

the previous section, the LD in this region was too extensive to directly pinpoint candidates (Fig. 5b), and we attempted to fine-map this QTL by generating a population of near-isogenic lines (NILs) (Methods). Based on the pattern of recombination and phenotype measurements for recombinant genotypes from BC_4F_2 and BC_4F_4 (Fig. 5c and Supplementary Tables 20–22), the QTL was fine-mapped to a 134-kb region

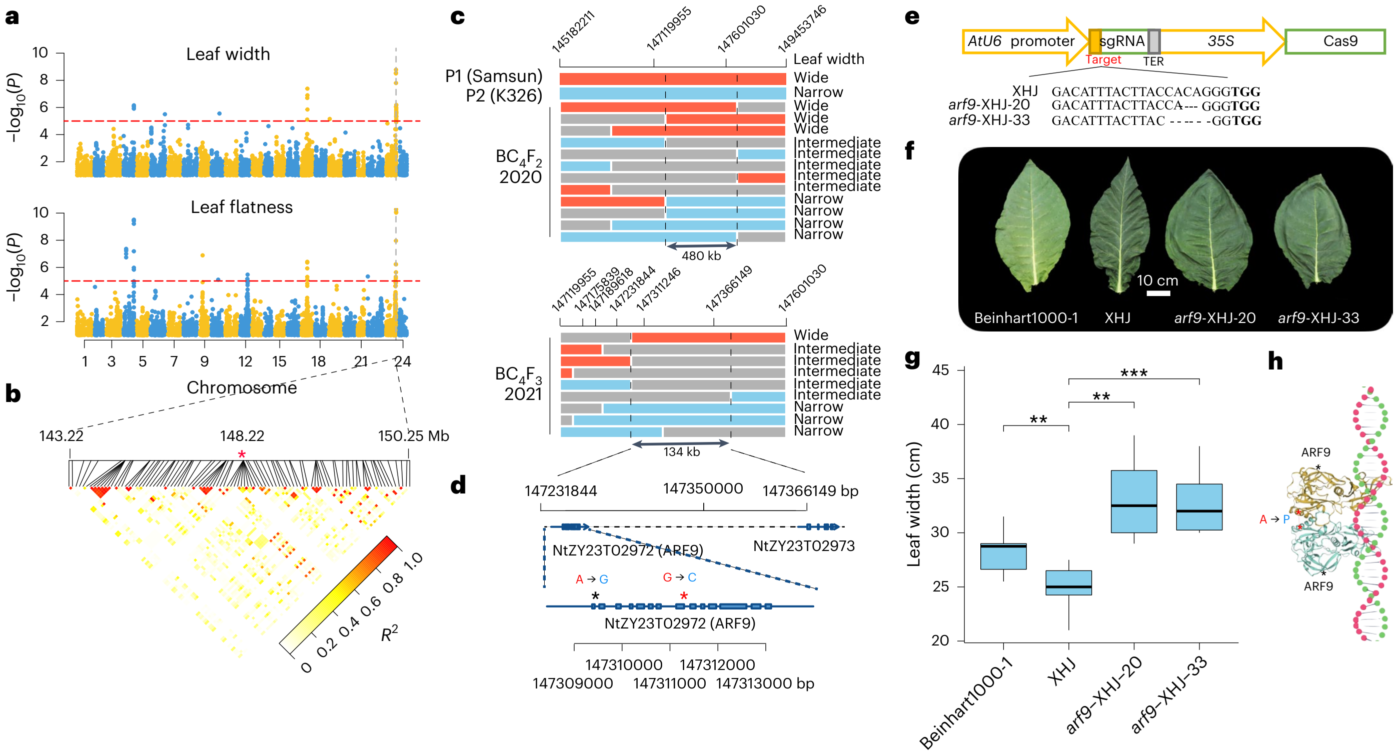

Fig. 5 | Fine-mapping and functional characterization of the QTL located at chr. 23:148211202, associated with LW variation. a, Association results for LF and LW. The red horizontal dashed lines indicate the Bonferroni-corrected genome-wide significance thresholds. **b**, The LD heatmap and literature review pinpointed *NtZY23G02972*, a homolog of *A. thaliana Arf9*, which regulates cell division activity, as a potential candidate underlying the association peak. **c**, Fine-mapping using NILs. Each bar represents a unique genotype in this QTL region. Red represents the genotype of the donor parent (P1; Samsum) with wider leaves, while blue highlights the genotype of the recurrent parent (P2; K326) with narrower leaves. The genotype classes of unique offspring from two generations, BC_4F_2 and BC_4F_3, are illustrated using colored bars, and the lines of origin are highlighted in red, blue and gray for P1 homozygous, P2 homozygous and heterozygous, respectively. **d**, Gene structure in the fine-mapped region and position of the two SNPs segregating in this population. A highly likely causal SNP is highlighted using red stars. **e**, Targeted sequences of *NtZY23T02972* (*Arf9*) in CRISPR–Cas9 knockout lines. **f**, Illustration of the LW of a control line (Beinhart 1000-1) and a targeted line before (XHJ) and after (*arf9*-XHJ-20, *arf9*-XHJ-33) CRISPR–Cas9 knockout. **g**, Boxplot of the LW of a control line (Beinhart 1000-1) and a targeted line before (XHJ) and after (*arf9*-XHJ-20, *arf9*-XHJ-33) CRISPR–Cas9 knockout. In the boxplots, the center line is the median, box limits are the first and third quartiles, and whiskers are the minimum and maximum. A two-sided *t*-test with a sample size of five was performed to generate the significances. $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$. **h**, A SNP altered the protein sequence of ARF from A > P, potentially affecting the formation of a homodimer to generate cooperative DNA binding. R^2, pairwise linkage disequilibrium.

between two markers, chr. 23:147231844 and chr. 23:147366149 (Fig. 5c). There were only two SNPs between the two parents (Fig. 5d) and one gene, *NtZY23G02972*, in this region. *NtZY23G02972* is a homolog of *Arabidopsis thaliana* Auxin Response Factor 9 (*Arf9*), which widely exists in a number of crop species but has not been functionally characterized. In *N. tabacum* var. XHJ, the LW of the *NtZY23G02972* CRISPR–Cas9 knockout line increased by 6 cm ($P = 2.74 \times 10^{-3}$) (Fig. 5e–g and Supplementary Fig. 14), suggesting that the nonfunctional allele of *NtZY23G02972* was associated with wider leaves. Previously, the structure of a homologous protein, ARF5, was characterized in *A. thaliana*, and ARF DNA-binding domains are known to form a homodimer to promote cooperative DNA binding, which is critical for in vivo ARF5 function[26]. Here we found that the second SNP on chr. 23:147311246, located in the eighth exon of *NtZY23G02972*, altered the translated amino acid sequence from alanine (Ala258) to proline (Pro258), and the first SNP was located in the first exon of *Arf9*, altering the amino acid sequence from glutamic acid (Glu26) to glycine (Gly26). This second SNP was located inside the functional domain, forming a homodimer (Fig. 5h), while the first SNP was at the proximal end of the protein, which is less likely to have an impact on protein activity. Taken together, it is highly likely that a causal variant at chr. 23:147311246 affects the formation of homodimers by changing the amino acid sequence from alanine (Ala258) to proline (Pro258) and alters the LW. In *A. thaliana*, *Arf9* is highly expressed in roots, and the transfer DNA insertion line does not exhibit an obvious auxin-related growth phenotype, possibly because of functional redundancy with other ARFs. Our results revealed a previously unrecognised role for *N. tabacum Arf9* in regulating leaf development and identified a likely causal candidate for the role of *Arf9* in leaf shape determination.

Signatures of positive selection and polygenic selection

When the germplasms were classified as landraces, introduced varieties (varieties developed from abroad) and local Chinese varieties (abbreviated as varieties hereafter) developed over a period of several hundred years, seven traits—LW, leaf length, FT, budding time, reducing sugar, total sugar and RBSH—continuously increased or decreased during the process of breeding (Fig. 6a–h). In total, 46.42% of the QTLs (highlighted using black or red stars in Fig. 6i) showed an increase in allele frequency during the process of breeding. However, none of the alleles reached fixation (frequency >0.95) and only 17.00% of the QTLs had allele frequencies >0.8, indicating that there is great potential for future genetic improvement. Simulation analysis revealed that the frequency of 42.86% of the randomly selected alleles increased because of stochasticity, suggesting that these QTLs may not be under strong positive selection (Supplementary Note). However, three QTLs (10.71%) displayed lower allele frequencies in landraces than in introduced

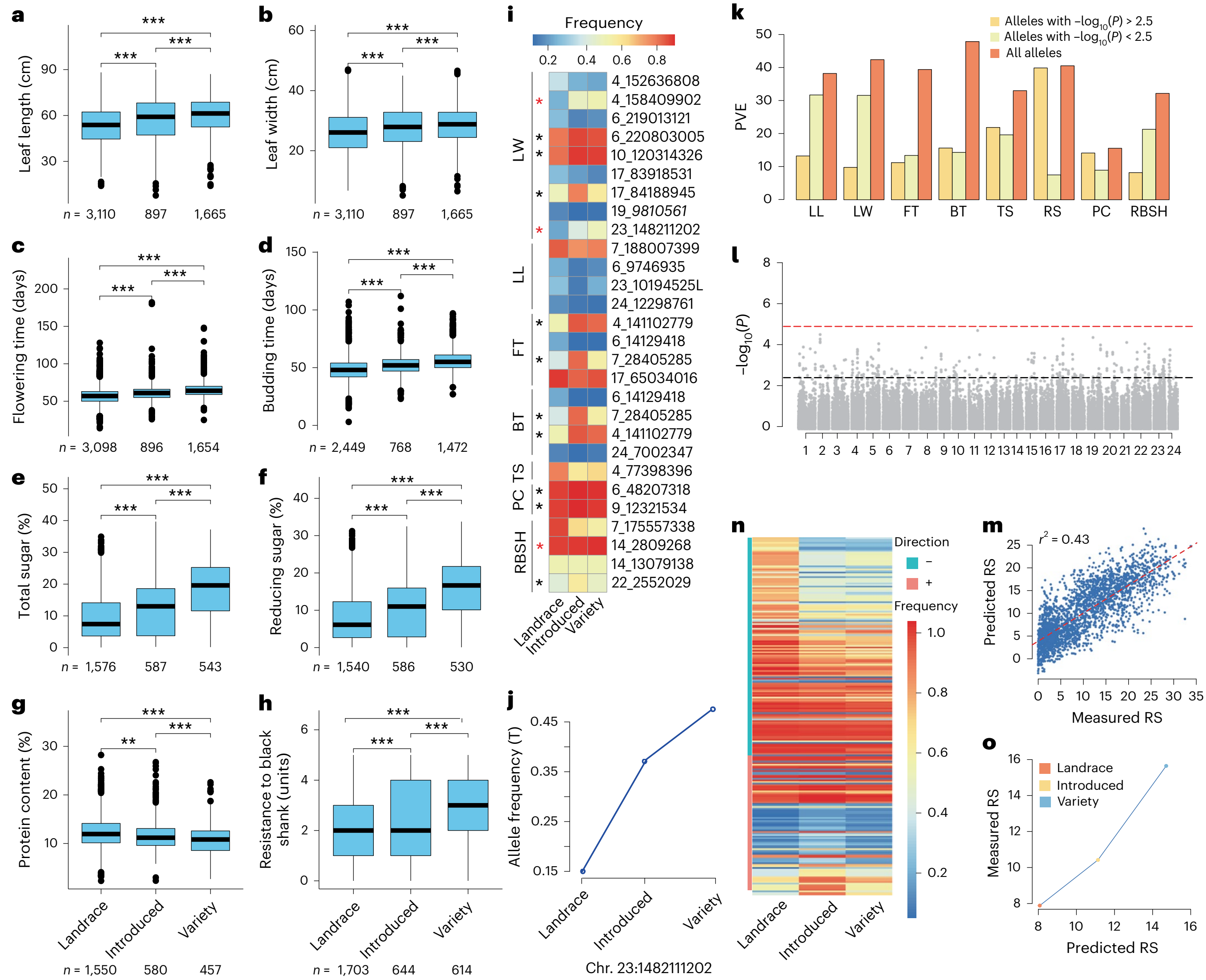

Fig. 6 | Selection signatures during selective breeding. a–h, Phenotype distribution of eight traits among different categories of materials (landraces, introduced varieties and varieties) developed during the process of breeding: leaf length (**a**), LW (**b**), FT (**c**), budding time (**d**), total sugar (**e**), reducing sugar (**f**), protein content (**g**) and RBSH (**h**). In the boxplots, the center line is the median, box limits are the first and third quartiles and whiskers are the minimum and maximum. Two-sided *t*-tests were performed to generate the significances. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. Exact *P* values from left to right are: 4.10×10^{-12}, 2.22×10^{-16} and 1.50×10^{-9} (**a**); 1.00×10^{-8}, 2.22×10^{-16} and 8.30×10^{-5} (**b**); 2.02×10^{-16}, 3.12×10^{-18} and 6.22×10^{-12} (**c**); 3.45×10^{-16}, 5.10×10^{-16} and 8.02×10^{-17} (**d**); 4.80×10^{-14}, 2.22×10^{-16} and 6.73×10^{-17} (**e**); 1.20×10^{-11}, 2.22×10^{-16} and 2.63×10^{-18} (**f**); 9.10×10^{-3}, 2.10×10^{-15} and 2.90×10^{-5} (**g**); and 5.00×10^{-4}, 2.20×10^{-15} and 1.30×10^{-12} (**h**). **i**, Frequency shifts of 28 QTLs associated with the variation in eight traits. **j**, Frequency shifts of the QTL associated with LW. **k**, The proportion of variance explained (PVE) by the alleles with $-\log_{10}(P)$ below or above 2.5 and all the alleles was estimated in a linear mixed model. **l**, Manhattan plot for the RS association analysis. Red horizontal dashed lines indicate the Bonferroni-corrected genome-wide significance thresholds. **m**, Scatterplot between predicted RS values based on 229 alleles with $-\log_{10}(P) > 2.5$ and measured alleles. **n**, Frequency shifts of the 229 alleles with $-\log_{10}(P) > 2.5$ according to the genome-wide association analysis of RS. **o**, Relationship between the mean predicted RS level based on 229 alleles with $-\log_{10}(P) > 2.5$ and the measured RS values for three categories of plant materials developed during the process of breeding. BT, budding time; LL, leaf length; PC, protein component; RS, reducing sugar; TS, total sugar.

varieties, followed by present-day varieties (red stars in Fig. 6i). For example, one QTL located at chr. 23:148211202 is associated with LW variation. The T allele at this locus increased the LW by 1.11 cm (Fig. 6j) ($P = 1.67 \times 10^{-9}$), and the frequency of the T allele continuously increased from 0.15 (landraces) to 0.38 (introduced varieties) and then to 0.47 (varieties) (Fig. 6j), suggesting that this allele is under positive selection.

In addition, polygenic scores were calculated using alleles with $-\log_{10}(P) > 2.5$ and $-\log_{10}(P) < 2.5$ using a mixed model with two random effects, representing the aggregated effects of minor-effect alleles and polygenic genetic background. Overall, these minor-effect alleles (0.26–0.29% of all SNPs) disproportionately accounted for a larger fraction (23.12–98.37%, median = 33.69%) (Fig. 6k) of the kinship heritability, suggesting an important contribution from the minor-effect alleles to the variation in these traits. In four of the eight cases, they explained more variance than did the remaining SNPs. Taking RS as an example, although no significant association was detected at the genome-wide significance threshold, the aggregated effects of 229 (0.23%) alleles with $-\log_{10}(P) > 2.5$ (Fig. 6l) explained 43.75% of the phenotypic variation and 95.56% of the kinship heritability (Fig. 6l,m). This was approximately five times greater than that for the SNPs with

$-\log_{10}(P) < 2.5$, as estimated by fitting a mixed linear model with two random effects. In total, 37.55% (86) of the alleles showed an increase in allele frequency from the landraces to the introduced varieties or varieties (Fig. 6n). Although the magnitude of the frequency increases was small (mean = 0.07) (Fig. 6n), they increased the level of reducing sugars from 8.07% to 14.71% from landraces to varieties (Fig. 6o), demonstrating the power of polygenic adaptation in response to artificial selection. Overall, these results demonstrate the power of leveraging GeneBank genomics for dissecting the genetic basis of complex trait variation and evolution during the process of selective breeding and provide a blueprint for future crop improvement.

Discussion

Subgenome gene expression dominance has been reported for several recent allopolyploids, such as strawberry[27], peanut[28], monkeyflower[8] and synthetic *B. napus*[29]. However, some allopolyploids exhibit even subgenome expression, including *Capsella bursa-pastoris*[30,31], *Trifolium repens*[32], *Arabidopsis kamachatica*[33], *Arabidopsis suecica*[27] and *Brachypodium hybridum*[34]. Our comparative genome and epigenomic analysis provided substantial evidence of chromosome rearrangement and subgenome gene expression divergence likely driven by epigenetic modifications for *N. tabacum* (Supplementary Note). We observed subgenome divergence in the genetic regulation of many complex traits and a fraction of the observed divergence could be attributed to genes biased toward one subgenome. Because subgenome divergence contributes to complex trait variation, artificial selection on traits whose regulation is biased toward one subgenome could drive subgenome transcriptomic and epigenomic divergence (Supplementary Note). We constructed a comprehensive genotype-to-phenotype map of this model plant species and demonstrated the power of this roadmap in plant functional genetics by fine-mapping one of the detected QTLs to a novel gene, *Arf9*, that regulates LW. Homologs of *Arf9* are widely present in a number of field crops, such as maize, rice and wheat, but have not been functionally characterized. It is worthwhile to evaluate the function and potential of *Arf9* in major crop molecular genetic research and breeding applications.

In summary, we presented chromosome-scale assemblies of *N. tabacum*, *N. sylvestris* and *N. tomentosiformis*, revealed genome rearrangements and subgenome transcriptomic and epigenomic divergence, and detected global genetic and phenotypic polymorphisms by sequencing and phenotyping an entire GeneBank collection of 5,196 *N. tabacum* germplasm resources. Our study provides insights into subgenome evolution and the genetic regulation of complex traits in polyploid species. The genome assemblies, extensive genotype and phenotypic datasets, marker–trait associations, and candidate genes presented in this study will serve as a community resource for accelerating future comparative genomics, plant functional genomics and crop improvement research.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-025-02126-0.

References

1. Sierro, N. et al. The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* **5**, 3833 (2014).
2. Sierro, N. et al. Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biol.* **14**, R60 (2013).
3. Leitch, I. J. et al. The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann. Bot.* **101**, 805–814 (2008).
4. Parisod, C. et al. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. *New Phytol.* **184**, 1003–1015 (2009).
5. Szadkowski, E. et al. The first meiosis of resynthesized *Brassica napus*, a genome blender. *New Phytol.* **186**, 102–112 (2010).
6. Xiong, Z. et al. Chromosome inheritance and meiotic stability in allopolyploid *Brassica napus*. *G3 (Bethesda)* **11**, jkaa011 (2021).
7. Vicient, C. M. & Casacuberta, J. M. Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* **120**, 195–207 (2017).
8. Edger, P. P. et al. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* **29**, 2150–2167 (2017).
9. Li, N. et al. DNA methylation repatterning accompanying hybridization, whole genome doubling and homoeolog exchange in nascent segmental rice allotetraploids. *New Phytol.* **223**, 979–992 (2019).
10. Johnson, C. S. Review: Tobacco: Production, Chemistry and Technology by D Layten Davis, Mark T Nielsen. *Q. Rev. Biol.* **77**, 66 (2002).
11. Lan, T. et al. Mapping of quantitative trait loci conferring resistance to bacterial wilt in tobacco (*Nicotiana tabacum* L.). *Plant Breed.* **133**, 672–677 (2014).
12. Yuan, G. et al. Development of a MAGIC population and high-resolution quantitative trait mapping for nicotine content in tobacco. *Front. Plant Sci.* **13**, 1086950 (2022).
13. Liu, Y. et al. Identification of QTLs associated with agronomic traits in tobacco via a biparental population and an eight-way MAGIC population. *Front. Plant Sci.* **13**, 878267 (2022).
14. Shi, R., Jin, J., Nifong, J. M., Shew, D. & Lewis, R. S. Homoeologous chromosome exchange explains the creation of a QTL affecting soil-borne pathogen resistance in tobacco. *Plant Biotechnol. J.* **20**, 47–58 (2022).
15. Sallaud, C. et al. Characterization of two genes for the biosynthesis of the labdane diterpene *Z*-abienol in tobacco (*Nicotiana tabacum*) glandular trichomes. *Plant J.* **72**, 1–17 (2012).
16. Edwards, K. D. et al. A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics* **18**, 448 (2017).
17. Schulthess, A. W. et al. Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. *Nat. Genet.* **54**, 1544–1552 (2022).
18. Milner, S. G. et al. Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* **51**, 319–326 (2019).
19. Nagaki, K., Shibata, F., Kanatani, A., Kashihara, K. & Murata, M. Isolation of centromeric-tandem repetitive DNA sequences by chromatin affinity purification using a HaloTag7-fused centromere-specific histone H3 in tobacco. *Plant Cell Rep.* **31**, 771–779 (2012).
20. Nagaki, K. et al. Coexistence of NtCENH3 and two retrotransposons in tobacco centromeres. *Chromosome Res.* **19**, 591–605 (2011).
21. Nagaki, K., Kashihara, K. & Murata, M. A centromeric DNA sequence colocalized with a centromere-specific histone H3 in tobacco. *Chromosoma* **118**, 249–257 (2009).
22. Lin, Y. et al. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023).
23. Renny-Byfield, S. et al. Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Mol. Biol. Evol.* **28**, 2843–2854 (2011).
24. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).

25. Cameron, D. R. & Moav, R. M. Inheritance in *Nicotiana tabacum* XXVII. Pollen killer, an alien genetic locus inducing abortion of microspores not carrying it. *Genetics* **42**, 326–335 (1957).
26. Boer, D. R. et al. Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. *Cell* **156**, 577–589 (2014).
27. Burns, R. et al. Gradual evolution of allopolyploidy in *Arabidopsis suecica*. *Nat. Ecol. Evol.* **5**, 1367–1381 (2021).
28. Zhuang, W. et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* **51**, 865–876 (2019).
29. Bird, K. A. et al. Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid *Brassica napus*. *New Phytol.* **230**, 354–371 (2021).
30. Kasianov, A. S. et al. High-quality genome assembly of *Capsella bursa-pastoris* reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. *Plant J.* **91**, 278–291 (2017).
31. Douglas, G. M. et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proc. Natl Acad. Sci. USA* **112**, 2806–2811 (2015).
32. Griffiths, A. G. et al. Breaking free: the genomics of allopolyploidy-facilitated niche expansion in white clover. *Plant Cell* **31**, 1466–1487 (2019).
33. Akama, S., Shimizu-Inatsugi, R., Shimizu, K. K. & Sese, J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid *Arabidopsis*. *Nucleic Acids Res.* **42**, e46 (2014).
34. Gordon, S. P. et al. Gradual polyploid genome evolution revealed by pan-genomic analysis of *Brachypodium hybridum* and its diploid progenitors. *Nat. Commun.* **11**, 3670 (2020).

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

[1]Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao, China. [2]Key Laboratory for Bio-Resource and Eco-Environment of Ministry of Education & Sichuan Zoige Alpine Wetland Ecosystem National Observation and Research Station, College of Life Science, Sichuan University, Chengdu, China. [3]State Key Laboratory of Plant Diversity and Specialty Crops, Institute of Botany, Chinese Academy of Sciences, Beijing, China. [4]China National Botanical Garden, Beijing, China. [5]University of Chinese Academy of Sciences, Beijing, China. [6]Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden. [7]CAAS-IRRI Joint Laboratory for Genomics-assisted Germplasm Enhancement, Agricultural Genomics Institute in Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [8]Strategic Innovation Platform, International Rice Research Institute, Metro Manila, Philippines. [9]Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark. [10]Beijing Life Science Academy, Beijing, China. [11]These authors contributed equally: Yanjun Zan, Shuai Chen, Min Ren, Guoxiang Liu, Yutong Liu. ✉e-mail: zanyanjun@caas.cn; chycht@163.com; chenglirui@caas.cn; yangaiguo@caas.cn

Methods

Plant materials, DNA extraction and sequencing

Total genomic DNA was collected and extracted from fresh leaves of *N. tabacum* L. var. ZY300 using the CTAB method[35]. For PacBio long-read sequencing, 20-kb insertion SMARTbell libraries were constructed and sequenced on the PacBio Sequel II platform (Pacific Biosciences). For 10X Genomics sequencing, 1 ng of genomic DNA with a long sequence length (approximately 50 kb) was partitioned by a microfluidic chip on the Chromium platform, and 16-bp barcodes were introduced into droplets. For Illumina short-read sequencing, paired-end libraries with insert sizes of 350 bp were constructed and sequenced on the HiSeq PE150 platform. For Hi-C sequencing, DpnII-digested and cross-linked DNA was labeled with biotin and proximity-ligated to form chimeric junctions. Biotin-labeled samples were captured and sheared into 350-bp fragments. After terminal repair, A addition, joint connection and library construction, Illumina sequencing was performed on the HiSeq PE150 platform. For Bionano sequencing, high-molecular weight DNA with a fragment distribution >150 kb was isolated using Bionano sample preparation kits (Bionano Genomics). Genomic DNA was labeled using Direct Label Enzyme (DLE-1) and stained following Bionano protocols. Then, the labeled DNA was loaded into a nanochannel (Saphyr Chip, Bionano Genomics) and imaged using the Saphyr system (Bionano Genomics) following the Saphyr System User Guide.

Low-quality paired reads (reads with ≥10% unidentified nucleotides (N); >10 nucleotides aligned to the adapter, >20% bases having a phred quality score <5 and putative polymerase chain reaction (PCR) duplicates generated during the library construction process), which resulted mainly from base-calling duplicates and adapter contamination, were removed. In total, these steps yielded 1.79 Tb of high-quality data for chromosome-scale scaffold analysis.

Estimation of genome size using *k*-mer and flow cytometry analysis

Jellyfish v.2.1.4 (ref. 36) was used to count the depth distribution of *k*-mer = 17–31. The spectrum of *k*-mer was subsequently fitted into findGSE[37] (v.1.94r) to estimate the genome size and heterozygosity. Flow cytometry analysis was performed according to a modified procedure reported in ref. 38. Tomato (R1: diploid, 900 M; R2: tetraploid, 1.8 G; R5: octoploid, 3.6 G) was used as an internal standard. Fifty milligrams of fresh leaf of *N. tabacum* (R6) and the standard were placed on ice in sterile 35- × 10-mm plastic Petri dishes. The tissues were chopped into pieces of about 1 mm, and soaked in a solution containing 50 mM KCl, 10 mM MgSO$_4$ 7H$_2$O, 3 mM dithiothreitol, 5 mM Hepes and 0.25% Triton X-100 at pH 8.0. Suspended nuclei were filtered with 30-μm nylon mesh. Propidium iodide (50 μg ml^{-1}) and DNase-free RNase (50 μg ml^{-1}) were used for staining. After filtration, samples were incubated for 30 min at 37 °C before being analyzed by a BD FACScalibur flow cytometer. The mean fluorescence intensity of the sample (*Is*) and the standards (*Ick*) was measured, and the genome size (*F*) was estimated according to the formula $F = Is/Ick \times n$, where *n* is the genome size of the standards.

Genome assembly, scaffolding and evaluations

For initial genome assembly of *N. tabacum*, PacBio long-read data were self-corrected to generate preassembled reads and assembled by the overlap-layout consensus algorithm[39] using Falcon[40] (v.1.0) with the following parameters 'overlap_filtering_setting = --max_diff 500--max_cov 500--min_cov 3--n_core 24--bestn 10'. The assembled contigs were further polished with Illumina short reads using the Pilon pipeline[41] (v.1.22) with default parameters. The polished contigs were anchored by 10X linked reads using fragScaff[42] (v.140324) with '-maxCore 200 -fs1 -m 3000 -q 30 -E 30000 -o 60000 -fs2 -C 5 -fs3 -j 1 -u 3'. For initial genome assembly of the *N. sylvestris* and *N. tomentosiformis*, a hybrid assembly approach combing HiFi reads from PacBio Revio, ONT ultralong-read data and Hi-C data were performed using hifiasm[43] assembler. Hi-C data were used to further improve the quality of the assemblies with HiC-Pro software[44] (v.2.10.0). Placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were adjusted manually using Juicebox[45] (v.1.11.08) to generate a final assembly. Pseudo-chromosomes for *N. tabacum* were named by mapping simple sequence repeats markers from a widely used linkage map[46] to unify the physical map and linkage map. Pseudo-chromosomes for *N. sylvestris* and *N. tomentosiformis* were named after their corresponding homologous chromosomes in the *N. tabacum* genome. The quality of the assembly was assessed in three ways. First, we mapped the Illumina short reads back to the assembled genome using Burrows–Wheeler aligner, Bwa[47] (v.0.7.17), mapping rates were summarized using SAMtools[48] (v.1.9) and number of SNPs were called using BCFtools[49] (v.1.17-50-ga8249495). Second, Universal Single-Copy Orthologues (BUSCO, v.3.0.2) analysis[50] was performed with the *Solanales* database (https://busco-data.ezlab.org/v5/data/lineages/solanales_odb10.2020-08-05.tar.gz). Last, *k*-mer completeness and base pair correctness were evaluated using Merqury[51] (v.1.4.1). The genome assembly has been uploaded to the National center for Biotechnology Information (NCBI) and a reviewer link was made for the revision (https://dataview.ncbi.nlm.nih.gov/object/PRJNA940510?reviewer=dt823s4k3l6te0vgf8g0j171un). The raw sequencing data, together with the genome assemblies are available as described in the 'Data availability' section.

Repeat and gene annotation

Extensive de novo TE Annotator (EDTA v.1.9.3)[52], a pipeline combining homology-based and de novo prediction methods using LTRharvest[53], LTR FINDER[53], LTR retriever[54], TIR-learner[55], Helitron Scanner[56] and Repeat Modeler[57], was used to annotate TEs, estimate TE insertion time and generate a species-specific library for gene annotation. Structural annotation of genes was conducted through a combination of homology-based, transcriptome-based and ab initio-based methods. For homolog prediction, protein sequences of plants, including *N. tabacum* (https://solgenomics.net/ftp/genomes/Nicotiana_tabacum/edwards_et_al_2017/annotation/), *Solanum tuberosum* (GCF_000226075.1), *Solanum lycopersicum* (GCF_000188115.4), *Coffea canephora* (PRJEB4211_v1), *Gossypium hirsutum* (GCF_000987745.1), *A. thaliana* (GCA_000001735.1) and *Vitis vinifera* (http://jul2018-plants.ensembl.org/Vitis_vinifera/Info/Index), were used as queries to search against the ZY300 genome using GeneWise (v.2.4.1)[58]. For transcriptome-based gene prediction, trimmed RNA-seq reads from stems, roots, leaves, anthers, flowers and axillary buds were de novo assembled using Trinity (v.2.1.1)[59]. To further improve the annotation, RNA-seq reads from different tissues were aligned to the ZY300 genome using TopHat (v.2.0.11) with default parameters to identify exons and splice junctions. The alignment was used as input for Cufflinks (v.2.2.1)[60] for transcript assembly with default parameters. For ab initio prediction, we used Augustus[61], GlimmerHMM[62], SNAP[63], GeneID (v.1.4)[64] and Genscan[65] to predict gene structure. Finally, all predictions of gene models generated from these approaches were integrated into a consensus gene set using EVidenceModeler (v.1.1.170)[66]. After prediction, PASA[67] was used to update isoforms to gene models and to produce a final gff3 file with three rounds of iteration.

For functional annotation, predicted protein-coding genes were aligned to multiple public databases, including NR, Swiss-Prot, TrEMBL75, COG and KOG, using NCBI BLAST+ v.2.2.31 (ref. 68). Motifs and domains were annotated using InterProScan (release 5.32-71.0)[69]. Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways of predicted sequences were assigned by InterProScan and KEGG Automatic Annotation Server[69,70], respectively.

Subgenome partitioning and synteny analysis

Illumina short-read sequences from *N. sylvestris* (accession ID: ERR274529) and *N. tomentosiformis* (accession ID: ERR274543) were downloaded from the NCBI SRA database using SRA toolkit and aligned to the ZY300 reference genome using bwa-mem[47]. SAMtools[48] was used to count the number of reads mapped to the ZY300 genome from each of the two ancestral genomes, and the average read depth in 1-Mb windows was calculated using a customized R script. First, a preliminary ancestral origin assignment was made using the window-averaged read ratio. Second, for windows containing breakpoints, manual curation was performed in Integrative Genomics Viewer[71] to pinpoint the exact location. WGDI[72] was used to detect synteny blocks based on collinear genes, and the R package circlize[73] was used to illustrate blocks between subgenomes after merging large synteny blocks. The partitioned genome was further curated by performing genome alignment between *N. tabacum* and its ancestral genome using minimap2 (v.2.26r1175)[74].

Chromatin immunoprecipitation, sequencing and bioinformatic analysis to pinpoint centromeric region

The chromatin immunoprecipitation experiment was conducted following the protocol described in ref. 75. Briefly, ~3.0 g of leaves was collected from *N. tabacum* and fixed with 1% formaldehyde in 1× PBS buffer. After incubation on ice for 10 min under a vacuum, 2.5 ml of 2 M glycine was added for 5 min to terminate the cross-linking. The tissue was then washed four times with sterile water, dried and immediately frozen in liquid nitrogen. Nuclei were purified with Honda buffer (0.44 M sucrose, 1.25% Ficoll, 2.5% Dextran T40, 20 mM Hepes KOH pH 7.4, 0.5% Triton X-100, 10 mM $MgCl_2$, 2 complete tablets per 100 ml), and subsequently resuspended in 600 µl of RIPA buffer (10% PBS 10X, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 5 mM dithiothreitol). DNA was released from the nuclei and fragmented using sonication at 4 °C for 5 min (30 s on and 30 s off) twice. After sonication, a 50-µl sample was used as DNA input. Then 900 µl of dilution buffer (16.7 mM Tris–HCl pH 8, 1.2 mM EDTA, 1.1% Triton X-100, 167 mM NaCl) was added to the remaining sample, which was incubated with 50 µl of Pierce Protein G Magnetic Beads (Thermo Fisher Scientific) that were immunoprecipitated with anti-CENH3 antibody (Abcam: ab72001). Immunoprecipitation was conducted at 4 °C for 4 h. Thereafter, the magnetic beads were washed with high-salt buffer (500 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl pH 8), low-salt buffer (150 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl pH 8) and TE buffer (10 mM Tris–HCl pH 8, 1 mM EDTA) twice for 5 min each time. The chromatin immunoprecipitation complex was then eluted from the beads with elution buffer (1% SDS, 0.1 M $NaHCO_3$). Reverse cross-linking was done by boiling the beads at 65 °C for 12 h in the presence of 10% SDS followed by proteinase K treatment at 50 °C for 1 h. DNA was extracted using phenol–chloroform and precipitated by ethanol. The air-dried DNA was dissolved in 20 µl of PCR-grade water (Roche) for sequencing library construction using NEXTFLEX ChIP-Seq Library Prep Kit (PerkinElmer) and sequenced by the Nova seq 6000 Illumina platform by Novogene. Three replicates were performed and sequenced to approximately 10X.

Raw reads were first trimmed using fastp[76] with default parameters and subsequently aligned to the ZY300 reference genome using Bwa[47]; SAMtoools[48] and Sambamba[77] were used to filter multimapped reads and eliminate PCR duplicates. Coverage information was then obtained using deepTools[78]. In addition, we searched for typical *N. tabacum* centromeres repeats obtained from CENH3 ChIP-seq in previous studies in the assembled genome[19–21] and performed de novo centromere prediction based on tandem repeat monomers using quarTeT[22].

Comparative transcriptome and epigenome analysis

Fresh leaves at the five-leaf stage were sampled for the two progenitors and *N. tabacum*. Three replicates for each plant were sent to Novogene for RNA-seq and bisulfate-seq. Raw RNA-seq data reads were first trimmed using fastp[76] with the default parameters and subsequently aligned to the ZY300 reference genome using HISAT2 (ref. 79). The expression level was quantified as read counts using StringTie[80]. To compare gene expression between *N. tabacum* and the two ancestral species, we partitioned the read count to the S and T subgenomes and constructed corresponding specific normalization factors to account for library size variation[81]. Scaled count matrices for each subgenome were further processed using a multifactor design to account for confounding factors (~subgenome+ replicates). To compare homologous gene expression between subgenomes, and between ancestors, expression analysis was then further restricted to 28,143 1:1 unique homologous gene pairs. We combined the read counts of homologous gene pairs from *N. sylvestris* and *N. tomentosiformis* into a single count matrix, and normalized confounding factors using a multifactor design (~ancestor). Wilcox test was used to calculate *P* values for all the comparisons.

Raw bisulfate-seq data reads were first trimmed using fastp[76] with default parameters and subsequently aligned to the ZY300 reference genome using Bismark[82]. CG, CHG and CHH methylation information was extracted using the BISMARK_METHYLATION_EXTRACTOR function implemented in Bismark[82]. The methylation level of each cytosine was calculated as the number of methylated cytosine reads divided by read depth. Sites covered by fewer than three mapped reads were filtered. To evaluate gene methylation levels, we computed the mean methylation level across all cytosine sites in the gene body and its adjacent 2-kb regions using methylKit[83].

Genotyping by sequencing, read mapping and variant calling

DNA extraction was performed using CTAB methods[35]. After quality control using Nanodrop and Qubit, DNA samples were digested with NlaIII + MseI, barcoded and purified using AMPure XP beads. Quality-controlled libraries were pooled and sequenced on the Illumina NovaSeq platform. Raw reads were trimmed using fastp[76] following the default parameters and subsequently aligned to the ZY300 reference genome using Bwa[47]; SAMtoools[48] mpileup was used for variant calling. The called SNPs were initially filtered for read depth, call rate and minor allele frequency (MAF) with vcftools[84] on the basis of an average read depth >2 and MAF > 0.03, individual missing rate <0.5 and site missing rate <0.5. The filtered genotypes were then imputed using Beagle[85] with default parameters. The imputed SNPs were further filtered for MAF > 0.03, and 95,308 SNPs remained for downstream analysis. Detailed information on the number of sequenced reads, mapping quality, genome coverage and sequencing coverage are summarized in Supplementary Fig. 16. To assess the genotype quality, we genotyped 42 individuals using whole-genome resequencing. SNP calling accuracy and imputation accuracy were estimated by comparing genotypes called from whole-genome resequencing with those obtained from genotype-by-sequencing before and after imputation. The mean accuracy of SNP calling was 0.958 (median = 0.976) before imputation, indicating high confidence in SNP calling using genotype-by-sequencing data. After imputation, the averaged accuracy was 0.941 (median = 0.976), and accuracy across a wide range of MAF was high (Supplementary Fig. 16), suggesting imputation by Beagle should not bias the accuracy of our results.

Analysis of population differentiation and genetic similarity

To further assess the relatedness between individuals, principal component analysis was performed using the IBS genetic distance matrix calculated in PLINK (v.1.90)[86] and the princomp function in the R base package. After grouping samples, VCFtools (v.0.1.16)[84] was used to calculate F_{ST} among major tobacco types and genetic clusters.

Collection of the germplasm and phenotyping

Systematic collection of the germplasm was started in the early 1950s, where landraces were collected from farmers across China. The majority of these germplasms were sun-cured tobacco types, which are well-adapted to the local environments across China. All collected germplasm resources are preserved in the GeneBank of Tobacco hosted at the Tobacco Research Institute, Chinese Academy of Agriculture

Sciences. Since then, an exchange of germplasm has been constantly performed with countries all over the world and imported materials have been crossed with landraces to introduce desired traits into local varieties. Detailed information on the origin and metadata of each sample are available in Supplementary Table 23.

In 2007, we began to phenotype the entire GeneBank hosted at the Tobacco Research Institute, Chinese Academy of Agriculture Sciences for 43 traits. Each year, around 500 plants were selected and planted in Qingdao, China (120.45° E, 36.38° N) under a randomized complete block design with two replicates. In each trial, one replication of each line includes two rows, each with ten replicates of the same genotype. Each row is 10 m in length with a row spacing of 1.2 m and a plant distance of 0.5 m. Each year, 500 lines were selected from the Germplasm bank for phenotyping and reviving old seeds. Because of disease and disruption by wild animals, the number of logged phenotypes from each year has varied between 412 and 440 since 2007. Owing to the unexpected COVID interruption, no phenotypes were obtained in 2019 and 2020. Each year, we excluded lines with severe disease or that were destroyed by animals, resulting in 412–440 lines with measured phenotypes from each year and 5,370 lines in total.

Detailed descriptions of trait measurements were mentioned in previous studies[12,13,87]. Briefly, traits related to flowering time were obtained by individually counting the days to budding and flowering. Plant architecture traits were manually measured for each plant using a ruler and angle ruler. Two mature middle leaves from each plant were harvested and measured for leaf morphological traits and metabolic traits. All phenotypes were measured immediately before the first flower blooms and averaged in each replicated row. For each year, averaged measurements among ten replicates were adjusted using a linear model, fitting blocks and rows as fixed effects, to calibrate the spatial variation, generating the phenotype records logged in the National Germplasm Database. For GWAS, we calibrated the yearly environmental effects for continuously distributed traits using a linear model fitting year as a fixed effect and extracted residuals as phenotypes (residuals are available in Supplementary Table 23)

Genome-wide association analysis

GWAS was conducted using a linear mixed model implemented in the mlma module of GCTA[88]. A subsequent conditional analysis implemented in the cojo module of GCTA[88] was performed to screen for independent association signals. Because LD was extensive in this population, assuming that all tested markers were statistically independent and deriving a Bonferroni-corrected significance threshold would have been too conservative. Therefore, we estimated the effective number of independent markers (*Me*)[89] and derived a less-conservative threshold following 0.05/*Me* (1.0×10^{-5} equivalent to $-\log_{10}(P) = 5$).

Polygenic selection analysis and genomic prediction

We predicted the aggregated effects of all the SNPs with a $-\log_{10}(P) > 2.5$ using the following model.

$$\mathbf{Y} = X\beta + Z\mathbf{u} + e \qquad (1)$$

where **Y** is a column vector of length *n* containing phenotype measurements. *X* is a matrix of *n* rows and one column of 1, representing the population mean. **u** is a random effect vector of the polygenic effects (polygenic scores) representing the aggregated effects of all the SNPs for *n* individuals. *Z* is the corresponding design matrix obtained from Cholesky decomposition of the kinship matrix *G*, estimated on the basis of the markers with a $-\log_{10}(P) > 2.5$ using the A.mat function in the R package rrBLUP[90]. The *Z* matrix satisfies $ZZ' = G$; therefore, $u \sim N(0, I\sigma_g^2)$. *e* is the residual variance with $e \sim N\left(0, I\sigma_e^2\right)$. The proportion of variance explained by these SNPs was estimated as $\sigma_g^2/(\sigma_g^2 + \sigma_e^2)$. Model fit was assessed in the R package rrBLUP[90].

Fine-mapping and experimental validation of candidate genes

A population of NILs was developed by marker-assisted backcrossing using Samsun (P1, T/T at chr. 23:148211202) as the wide-leaved donor parent and K326 (P2, C/C at chr. 23:148211202) as the narrow-leaved recurrent parent. First, a total of 1,694 BC_4F_2 individuals from the K326 × Samsun population were genotyped using five markers from chr. 23:145182211 and chr. 23:149453746 (Supplementary Table 20), selected recombinants were then self-pollinated to create BC_4F_3 lines for further evaluation. This narrowed the QTL to a region between chr. 23:147119955 and chr. 23:147360000. An additional 3,017 BC_4F_3 individuals derived from the population were further screened using seven SNP markers in this narrow region (Supplementary Table 20), and the BC_4F_4 lines of recombinants were phenotyped for further fine-mapping. Two small guide RNAs (Fig. 5e, synthesis by Beijing Genomics Institute) were designed to target the candidate gene *NtZY23T02972* based on the assembled ZY300 genome. The vectors were constructed and transformed into the receptor *N. tabacum* L var. *XHJ*. The genotype of gene-edited lines was identified by PCR amplification and Sanger sequencing. For fine-mapping, LW was recorded by measuring at least 15 plants per line during flowering in Sanya (108.56° E, 18.09° N), China. For transgenic experiments, phenotypes of knockout lines and the wild-type were investigated in Sanya (108.56° E, 18.09° N), China. Each experiment had two replicates with at least five independent plants per replicate. The mean LW of at least five plants per replicate was used for further analyses.

A coreset of seeds is publicly available for the community

To make these resources accessible to a wide range of researchers, we selected 310 accessions to cover the majority of the genetic and phenotypic diversity (Supplementary Fig. 15 and Supplementary Table 24), and have made their seeds available for the research community (see https://www.cgris.net/search and https://yanjunzan.github.io/Resources/ for detailed ordering information, commercial use is strictly prohibited). Our intention is for this collection to remain actively curated as ever more genomic data are produced and a wide range of phenotypic data are generated not only by us, but also by the community.

Statistical analysis

Details on all statistical analyses used in this paper, including the statistical tests used, the number of replicates and precision measures, are indicated in the corresponding section and figure legends. Statistical analysis of replicate data was performed using appropriate strategies in R (v.4.4.1).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The genome assemblies of *Nicotiana sylvestris*, *Nicotiana tomentosiformis* and *Nicotiana tabacum* were deposited to European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) with accession codes PRJEB85697, PRJEB85577 and PRJEB85578, respectively. All raw sequences were uploaded to NCBI with accession codes PRJNA940510, PRJNA936601, PRJNA1074506, PRJNA1074481 and PRJNA1074687. Genome annotations are available via Figshare at https://doi.org/10.6084/m9.figshare.25139579.v1 (ref. 91), https://doi.org/10.6084/m9.figshare.25139468.v1 (ref. 92) and https://doi.org/10.6084/m9.figshare.25139474.v1 (ref. 93). All phenotype measurements and metadata are included in the Supplementary Tables.

Code availability

Analysis codes are available via GitHub at https://github.com/yanjunzan/N_tabacum_Genome_paper and via Zenodo at https://doi.org/10.5281/zenodo.14807329 (ref. 94).

References

35. Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320–2325 (2006).
36. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
37. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and *Arabidopsis* using *k*-mer frequencies. *Bioinformatics* **34**, 550–557 (2018).
38. Zhang, J. et al. Genome size variation in three *Saccharum* species. *Euphytica* **185**, 511–519 (2012).
39. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
40. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
41. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
42. Adey, A. et al. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).
43. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
44. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
45. Robinson, J. T. et al. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* **6**, 256–258.e1 (2018).
46. Bindler, G. et al. A high density genetic map of tobacco (*Nicotiana tabacum* L.) obtained from large scale microsatellite marker development. *Theor. Appl. Genet.* **123**, 219–230 (2011).
47. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
48. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
49. Narasimhan, V. et al. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).
50. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
51. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
52. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
53. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
54. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
55. Su, W., Gu, X. & Peterson, T. TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Mol. Plant* **12**, 447–460 (2019).
56. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl Acad. Sci. USA* **111**, 10263–10268 (2014).
57. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
58. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
59. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
60. Ghosh, S. & Chan, C.-K. K. Analysis of RNA-Seq data using TopHat and Cufflinks. *Methods Mol. Biol.* **1374**, 339–361 (2016).
61. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
62. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
63. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
64. Guigó, R. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comput. Biol.* **5**, 681–702 (1998).
65. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
66. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
67. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
68. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
69. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
70. Aoki-Kinoshita, K. F. & Kanehisa, M. in *Comparative Genomics* Vol. 2 (ed. Bergman, N. H.) 71–92 (Humana, 2007).
71. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* **14**, 178–192 (2013).
72. Sun, P. et al. WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).
73. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
74. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
75. Dong, Y. et al. Regulatory diversification of INDEHISCENT in the *Capsella* genus directs variation in fruit morphology. *Curr. Biol.* **29**, 1038–1046.e4 (2019).
76. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
77. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
78. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
79. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

80. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
81. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
82. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
83. Akalin, A. et al. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
84. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
85. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
86. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
87. Liu, Y. et al. Screening of tobacco genotypes for *Phytophthora nicotianae* resistance. *J. Vis. Exp.* **182**, e63054 (2022).
88. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
89. Li, M.-X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant *p*-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).
90. Endelman, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **4**, 250–255 (2011).
91. Zan, Y. Genome annotation for *N.tabacum*. *Figshare* https://doi.org/10.6084/m9.figshare.25139579.v1 (2024).
92. Zan, Y. Genome annotation for *N.sylvestris*. *Figshare* https://doi.org/10.6084/m9.figshare.25139468.v1 (2024).
93. Zan, Y. Genome annotation for *N.tomentosiformis*. *Figshare* https://doi.org/10.6084/m9.figshare.25139474.v1 (2024).
94. Zan, Y. Analysis code for 'The genome and GeneBank genomics of allotetraploid *Nicotiana tabacum* provide insights into genome evolution and complex trait regulation'. *Zenodo* https://doi.org/10.5281/zenodo.14807328 (2025).

Acknowledgements

This research was funded by Taishan Young Scholar Program (to Y. Zan) and Distinguished Overseas Young Talents Program from Shandong province (grant no. 2024HWYQ-079 to Y. Zan), the National Science Foundation of China (grant no. 32200503 to Y. Zan, 32170227 and 32470227 to Y. Dong), Agricultural Science and Technology Innovation Program (grant no. ASTIP-TRIC01 to A.Y.) from Chinese Academy Agriculture Sciences, the Crop Germplasm Resources Protection and Utilization Special Grant from the Ministry of Agriculture and Rural Affairs of the People's Republic of China, and the National Science and Technology Resource Sharing Service Platform (grant no. NCGRC-2023-18 to M.R. and A.Y.). We thank W. Zhao and Y. Guo for their constructive comments to improve the paper. We thank S. Tian for his valuable input during the design and implementation of this study.

Author contributions

Y. Zan, A.Y., L.C. and Y.C. designed and supervised this study. Y. Zan, Y.H., Y. Zhang, Y.D., S.C., M.R., G.L. and Y. Liu collected the data and implemented the analysis. Y.H. and H.S. contributed to the statistical analysis. Y. Li, X.Z., Y.T., C.J., Z.L., D.L., Y.S., Z.X., L.W. and R.G. conducted the field experiments and phenotype measurements. Q.F. and Y.W. collected the germplasms. G.Y. and L.F. offered valuable input during the design and implementation of this study. Y.J. prepared the figures. Y. Zan, S.C. and L.C. wrote the paper. All authors read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02126-0.

Correspondence and requests for materials should be addressed to Yanjun Zan, Yong Chen, Lirui Cheng or Aiguo Yang.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

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

nature portfolio

Corresponding author(s): Aiguo Yang, Lirui Cheng, Yong Chen

Last updated by author(s): Nov 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequencing platforms used to generate data in the paper are : Pacbio SMAR, Revio, Oxford Nanopore, and Illumina Novaseq
Data analysis	Software used in this paper and corresponding version are Jellyfish v.2.1.4, findGSE v.1.94, Falcon v1.0, Pilon v1.22, fragScaff v140324, HiC-Pro v.2.10.0,Juicebox v1.11.08, Bwa v0.7.17, SAMtools v1.9, BCFtools v1.17, BUSCO v3.0.2, Merqury v1.4.1, hifiasm v0.16, EDTA v1.9.3, GeneWise v.2.4.1, TopHat v2.0.11, Augustus v3.4.0, GlimmerHMM v3.0.2, SNAP v2.4.0, GeneID v1.4, Genscan, EVidenceModeler v1.1.170, PASA v2.5.3,BLAST+v.2.2.31, InterProScan 5.32-71.0, SRA toolkit v 2.11.0, IGV v2.17.0, WGDI v0.6.5, minimap2 v2.26r1175, Sambamba v1.0.1, deepTools v3.2.1, HISAT2 v2.1.0, Stringtie v2.2.0, fastp v0.20.0, Bismark v0.24.2,methylKit v1.32.0,vcftools v0.1.16, beagle v5.4, PLINK v1.90,GCTA v1.94.1,rrBLUP 4.6.3, R v.4.4.1. Analysis code is available on Github as described in the method section (https://github.com/yanjunzan/N_tabacum_Genome_paper)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All sequence data generated in this study had been uploaded to NCBI, and accession codes were also provided in Data availability section. The genome assembly and all sequence data were uploaded to NCBI with accession code PRJNA940510, PRJNA936601, PRJNA1074506, PRJNA1074481, and PRJNA1074687. Genome annotations were available in figshare (https://doi.org/10.6084/m9.figshare.25139579.v1, https://doi.org/10.6084/m9.figshare.25139468.v1, https://doi.org/10.6084/m9.figshare.25237585, https://doi.org/10.6084/m9.figshare.25139474.v1). All phenotype measurements and meta data of samples are included in Supplementary Tables.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	not relevant to this study.
Reporting on race, ethnicity, or other socially relevant groupings	not relevant to this study.
Population characteristics	not relevant to this study.
Recruitment	not relevant to this study.
Ethics oversight	not relevant to this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For GWAS, we have collected 5196 tobacco germplasms and genotyped in this study. This is determined by the number of accessions we have collected
Data exclusions	No data exclusion
Replication	For CRISPR/Cas9 plants, RNA-Seq, bisulfate-seq, we used three replicates for each experiment. For filed phenotyping we used 10 replicates.
Randomization	All the plants were randomized when phenotyping at field. Each year, around 500 plants were selected and planted in Qingdao, China (120.45° E, 36.38° N) with a randomized complete block design with two replicates. In each trial, one replication of each line includes two rows, each with 10 replicates of the same genotype. Each row is 10 meters in length with a row spacing of 1.2 m and a plant distance of 0.5 m.At each year, averaged measurements among ten replicates were adjusted using a linear model, fitting blocks and rows as fixed effects, to calibrate the spatial variation, generating the phenotype records logged into the National Germplasm Database. For genome-wide association analysis (GWAS), we calibrated the yearly environmental effects for continuously distributed traits using a linear model fitting year as a fixed effect and extracted residuals as phenotypes
Blinding	we collected the data without any prior knowledge of the plants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
□	☒ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
□	☒ Plants

Methods

n/a	Involved in the study
□	☒ ChIP-seq
□	☒ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used	anti-CENH3 antibody (Abcam number: ab72001).
Validation	This is a commercial antibody that have bee validated by six publication. https://www.abcam.com/en-us/products/primary-antibodies/cenh3-antibody-ab72001

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	□ Public health
☒	□ National security
☒	□ Crops and/or livestock
☒	□ Ecosystems
☒	□ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	□ Demonstrate how to render a vaccine ineffective
☒	□ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	□ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	□ Increase transmissibility of a pathogen
☒	□ Alter the host range of a pathogen
☒	□ Enable evasion of diagnostic/detection modalities
☒	□ Enable the weaponization of a biological agent or toxin
☒	□ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe*

	the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	Accession number PRJNA940510
Files in database submission	SRR27908885: Replication 1 with treatment and control groups using paired-end sequencing. SRR27908884: Replication 2 with treatment and control groups using paired-end sequencing. SRR27908881: Replication 3 with treatment and control groups using paired-end sequencing.
Genome browser session (e.g. UCSC)	no longer applicable

Methodology

Replicates	3
Sequencing depth	10X
Antibodies	Anti-CENH3 (Abcam number: ab72001)
Peak calling parameters	Peaks were identified by comparing the depth of coverage in the BAM file between the treatment and control groups using the deepTools bamCompare function, with parameters set to --scaleFactorsMethod SES, --binSize 5000, and --operation log2.
Data quality	Raw reads were first trimmed using fastp with default parameters and subsequently aligned to the ZY300 reference genome using bwa. SAMtoools and Sambamba was used to filter multi-mapped reads and eliminate PCR duplicates. Coverage information was then obtained using deepTools.
Software	fastp, bwa SAMtools Sambamba deeptools.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Sample preparation listed in Methods.
Instrument	BD FACScalibur flow cytometer
Software	Flowjo v10.8.1
Cell population abundance	Sorting was not employed in this study.
Gating strategy	Gates were empirically defined based on desities of measured standards.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.