GenBank genomics highlight the genomic features, genetic diversity and regulation of morphological, metabolic and disease-resistance traits in *Nicotiana tabacum*

Yanjun Zan[1#], Shuai Chen[1#], Min Ren[1#], Guoxiang Liu[1#], Yutong Liu[1#], 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], Quanfu Feng[1], Yuanying Wang[1], Yong Chen[2]*, Lirui Cheng[1]*, Aiguo Yang[1]*

[1]Key Laboratory of Tobacco Improvement and Biotechnology, Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao, 266000, China

[2]China National Tobacco Corporation, Beijing, 100045, China

[#] These authors contributed equally

[*]Corresponding authors

Aiguo Yang

Email: yangaiguo@caas.cn

Lirui Cheng

Email: chenglirui@caas.cn

Yong Chen

Email: chenyong@tobacco.gov.cn

Abstract:

Nicotiana tabacum is a model organism in plant molecular and pathogenic research, and has significant potential in producing biofuels and active pharmaceutical compounds in synthetic biology. Challenged by the large allotetraploid genome, the genomic features, genetic diversity and genetic regulation of many complex traits remain unknown. In this study, we present the nearly complete chromosome-scale assembly of *N. tabacum*, and provide evidence that chromosome rearrangement between subgenomes and epigenetic remodeling are likely mechanisms in genome stabilization and subgenome coordination following polyploidization. By leveraging GenBank scale sequencing and phenotyping data, geography at the continent scale, rather than types named after curing crop practices, was found to be the most important correlate of genetic structure. Using 178 marker-trait associations detected in genome-wide association analysis, a reference genotype-to-phenotype map was built for 39 plant morphological, developmental, and disease-resistant traits. A novel gene, auxin response factor 9 (*Arf9*), associated with wider leaves after being knocked out was fine-mapped to a SNP. This point mutation alters the translated amino acid from Ala_{203} to Pro_{203}, likely preventing the formation of a homodimer in DNA binding. Our analysis also revealed signatures of positive selection and polygenic selection for multiple traits during the process of selective breeding. Overall, this study demonstrated the power of leveraging GenBank genomics to gain insights into the genomic features, genetic diversity, and regulation of complex traits in *N. tabacum*, laying a foundation for future research in plant functional genomics, crop breeding, and the production of bio-pharmaceuticals and biofuels.

Introduction

Nicotiana species have served as models for a series of cytogenetic and molecular genetic breakthroughs, including the discovery of protoplast fusion[1], introduction of foreign genetic material through interspecies hybridization[2,3], T-DNA mediated plant transformation[4], virus-induced genetic transformation[5], report of the first edible vaccine[6,7], and plant pathogen studies[8]. The rich diversity in plant architecture, leaf morphology, flowering time and metabolic traits, ease of genetic manipulation, and fairly high number of curated pathways in *Nicotiana tabacum*, highlighted the central importance of *N.tabacum* in plant development research and significant potential in the fields of synthetic biology [6,7].

Nicotiana tabacum (2n=4x=48, ~4.30Gb) is an interspecific hybrid derived from two progenitors, *Nicotiana sylvestris* (2n=2x=24, 2.59Gb) and *Nicotiana tomentosiformis* (2n=2x=24, 2.22 Gb) about 0.2 Mya ago[9–11]. Several attempts have been made to assemble the large allotetraploid genome, however, the latest assembly only included 81% of the genomic sequences and is very fragmented[9,11,12]. As a consequence of its large and repetitive genome, obtaining a chromosome-level assembly constituted a significant challenge, as there are difficulties not only in bridging highly repetitive repeat regions but also in computationally disentangling homologous sequences with high similarity from two progenitor species[9,11,12]. In absence of a high-quality genome assembly and nearly all genetic studies being performed in the private sectors, the genomic features, global genetic diversity as well genetic basis of many important traits in this model plant remains poorly understood. These factors highlighted the necessity for a GenBank-scale holistic investigation of the genetic diversity and regulation of complex traits that are highly relevant for future plant functional genomics, crop improvement, and plant synthetic biology research.

In this study, we presented the first complete chromosome-level assembly of the allotetraploid *Nicotiana tabacum* genome, the genetic and phenotypic data on an entire *Nicotiana tabacum* GenBank hosted at Chinese Academy of Agriculture Sciences. By utilizing this valuable resource, insights were gained on the genomic features, genetic diversity and genetic regulation of plant architecture, morphological, metabolic, as well as disease-resistant traits in this model species. A novel gene, auxin response factor 9 (*Arf9*), confirms wider leaf after

being knocked out was fine-mapped to a SNP altering the translated amino acid from Ala_{203} to Pro_{203}, which in turn prevented the formation of a homodimer in DNA binding. Results from our study will likely accelerate aspects of tobacco research that were hindered by the complexity of the polyploid genome and a lack of publicity for research driven by private sectors, with benefits in plant functional genomics, crop improvement research, and production of bio-pharmaceutical compounds and biofuel.

Results

Chromosome-scale assembly and genomic feature of allotetraploid Nicotiana tabacum genome

Based on k-mer analysis with 52X Illumina short reads and flow cytometry analysis, the *N. tabacum* genome was estimated to be 4.29-4.30 Gb, representing approximately a 10% downsizing from the sum of the two parental genomes (Fig 1A, Supplementary Fig S1). To overcome the challenge of assembling a large polyploid genome, we generated 52X Illumina short reads, 123.35 X Pacbio Sequel II reads, 124.93X 10X Genomic linked reads, 120.12 X Hi-C reads and 192.27 X Bionano reads from *Nicotiana tabacum L. var ZY300* and adopted a hybrid assembly approach (Materials and methods). The final assembly included 4.17 Gb (Fig 1B) sequences with contig N50 at 27.17 Mb and 96.98% of the sequences anchored to 24 pseudo-chromosomes (Supplementary Table 1, Supplementary Fig S2). This represented a 10-15 percent increase in assembled genome size, and a 10-fold increase in contig N50 length compared to the latest assembly (Supplementary Table 1) [9,11,12]. The continuity and completeness of the assembled genome was supported by a high mapping rate (99.72%) and genome coverage (99.47%) of Illumina short reads, 99.2% of the single-copy orthologs from the Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis and long terminal repeat (LTR) assembly index (LAI) scores (LAI=9.74).

The complete assembly of a complex allotetraploid presented the opportunity to study subgenome evolution and coordination following polyploidisation. 57.14% and 42.85% of the genome is partitioned to *N. sylvestris* (S) and *N. tomentosiformis* (T) with 12 homolog exchange events pinpointed to approximately 1kb resolution (Supplementary Table 2). This suggested that chromosomal rearrangements, produced significant changes in genome structure, which likely stabilized chromosome pairing after polyploidization. To evaluate

mechanisms contributing to genome downsizing and subgenome coordination, we annotated the transposable elements (TE) and homolog gene pairs, and quantified their sequence and expression divergence. Based on a de novo-constructed species-specific TE library, 82.75% of the genome was masked. Estimation of the TE insertion time showed that approximately 7.56% of the TE were inserted after the merge of two subgenomes happened 0.2 Mya ago (Fig 1C), suggesting that polyploidisation did not start a massive reactivation of TE. Overall sequence similarity of 35740 homologous gene pairs between subgenomes was high (median = 92%, Fig 1D), and expression divergence was low (median fold change estimated from 400 RNA-seq data is 0.98, Fig 1E, Supplementary Table 3). Yet, over a thousand homology genes with high sequence similarity displayed considerable expression divergence (median fold change = 3.92, Fig 1E, Supplementary Table 3, Supplementary Fig S3), indicating a likely role of epigenetic remodeling in subgenome gene expression coordination for a small proportion of homologous genes. A detailed summary of the TE annotation and gene annotation statistics are available in Supplementary tables 4-5.

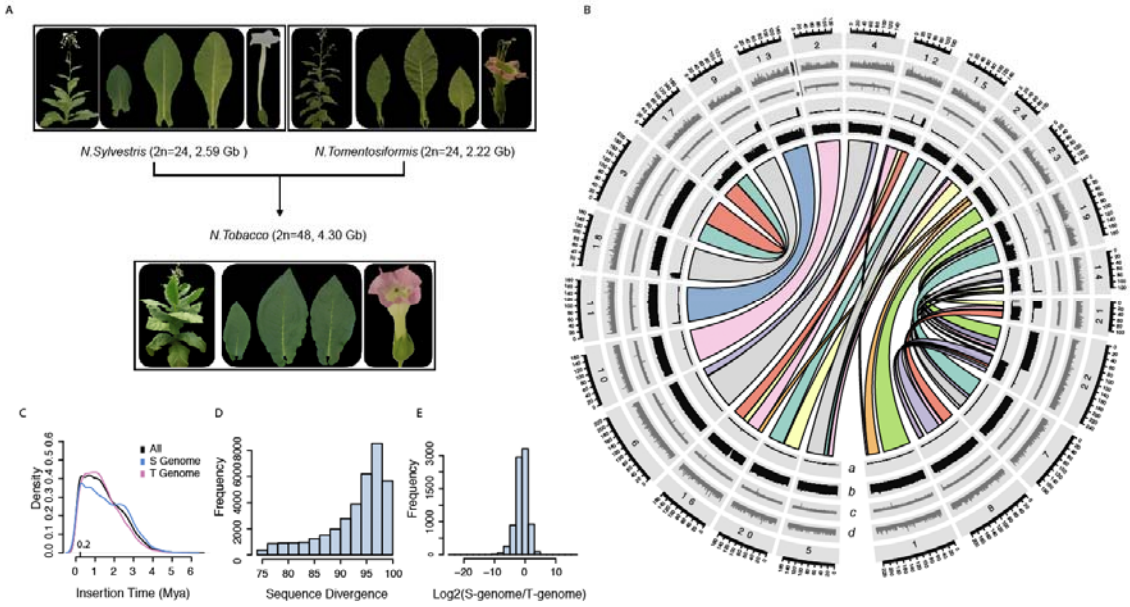

Fig 1. High-quality genome assembly and genomic features of *N. tabacum* genome. A) Phenotype diversity, and genome size variation among two diploid progenitors, *N. sylvestris*, *N. tomentosiformis* and one of the allotetraploid *N. tabacum* samples. **B)** Genomic features of *N. tabacum.* Tracks from innermost to outermost are sequence coverage obtained by mapping Illumina short reads from *N. tomentosiformis* (a), *N. sylvestris* (b), gene density (c), transposable element (TE) density (d). Synteny blocks between two subgenomes are

represented by coloured linkers across the centre of the plot. **C)** Estimated TE insertion time for TE from two subgenomes. Histogram of sequence divergence **D)** and expression divergence **E)** of 35740 protein-coding homologous gene pairs from two subgenomes.

GenBank genomics highlighted the diversity of a global collection of 5196 tobacco lines

To understand the global distribution of genetic variation and genetic structure, we genotyped an entire GenBank hosted at Chinese Academy of Agriculture Sciences using a genotype-by-sequencing approach (Materials and methods). This is one of the world largest collections of tobacco germplasm resources with 2582 sun-cured tobacco, 2152 flue-cured tobacco, 223 burley tobacco, 126 cigar tobacco and 113 oriental tobacco, amounting to 5196 germplasms in total (Supplementary Table 6). These five types of tobacco were named after their roles in agronomical practices (Supplementary Table 7), and characterized by distinctive characteristics in plant architecture, leaf morphology, flowering time and metabolic traits (Fig 2A, Supplementary Fig S4).

Using 95,308 SNPs covering 98% of the genomic bins in 1Mb resolution (Fig 2B), we found that geography at the continent scale was the most important correlate of genetic structure, and that genetic structure was not always matching the conventional types given after agronomical practices. For example, the first principle component (PC1) separated flue-cured tobacco from sun-cured tobacco to a large extent. However, the majority of samples located on the left side of Fig 2C were flue-cured tobacco originated from North America (NA) or derived materials with NA lineage, while plants located on the opposite side were sun-cured tobacco landraces collected from China (Fig 2C/2D). In addition, a latitude cline was observed along the second principal component (PC2). Sun-cured tobacco landrace from northern China (NC) with a significant shorter life cycle, colder growth temperature and longer day length was located in the upper right corner (Cluster NC, Fig 2C), while plants sampled from southern China (SC) with opposite climate signatures were clustered at the bottom of Fig 2C. According to historical records, burley tobacco originated from a single mutated variety, known as the White Burley, characterized by a chlorophyll deficiency phenotype underpinned by double homozygous recessive alleles at the Yellow burley 1 (*YB1*) and Yellow burley 2 (*YB2*) loci[13]. In consistent with this, burley tobacco samples were clustered together nearby NA group and showed the highest pairwise IBS value (median=0.88,

Fig 2E). Cigar and oriental tobacco samples were scattered among sun-cured and flue-cured tobacco, indicating a likely admixed breeding history (Fig 2C/D).

Pairwise IBS values for five tobacco types ranged from 0.75-0.88, suggesting that the gene pool to develop each type of tobacco was very narrow. Yet, the distribution of IBS for flue-cured, sun-cured and cigar tobacco was multimodal, indicating the co-occurrence of divergent gene pools likely due to introgression from foreign materials. Despite the remarkable phenotypic differentiation in plant morphology, flowering and metabolic traits (Fig 2A/ Supplementary Fig S4), genome-wide genetic differentiation among five types and three genetic groups was low, with pairwise F_{st} ranging from 0.04 to 0.23 (Fig 2F/G), and no selective sweep was detected (Data not show here). This is most likely due to the short and special selective breeding history, in which the selection history was less than a few hundred years and the targets of selection were balancing quality among many leaf characteristics.

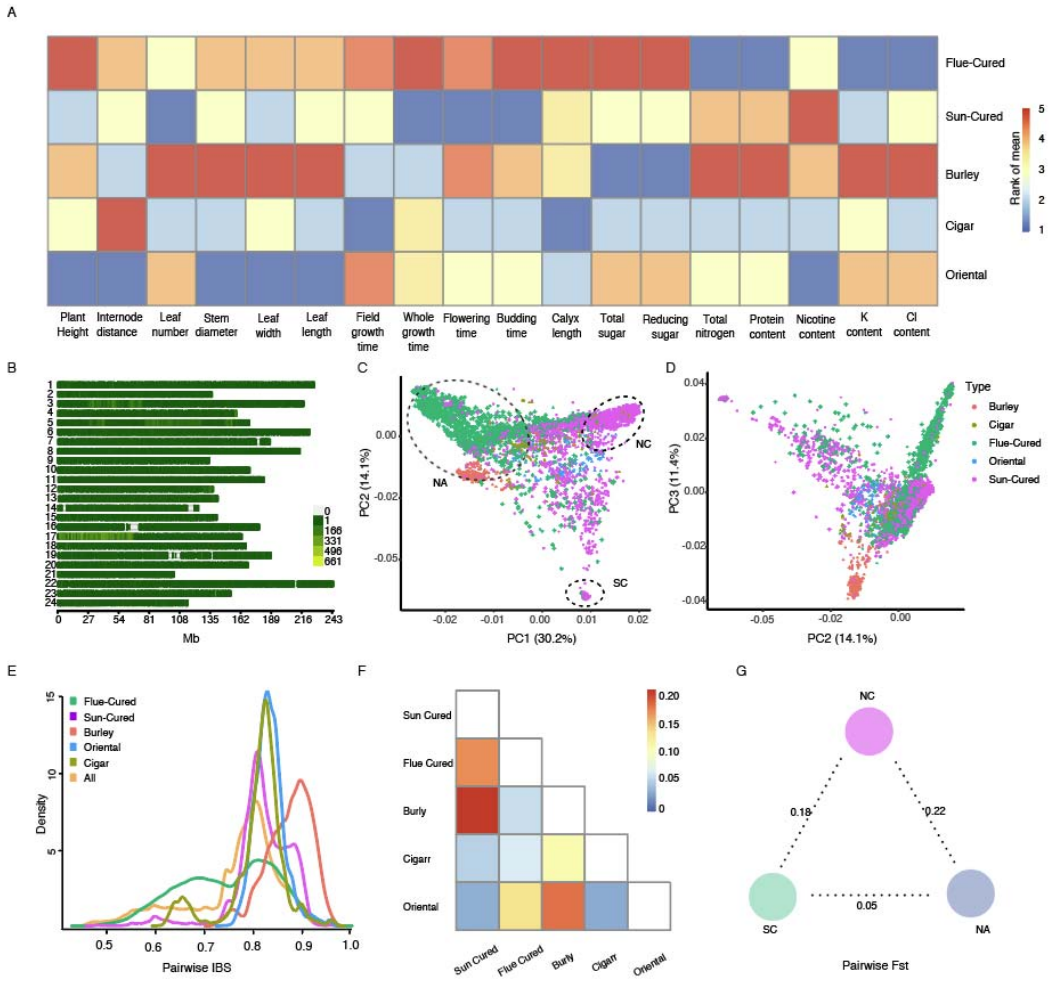

Fig 2. The genetic and phenotypic diversity of a global collection of 5,196 tobacco germplasm resources. A). A heatmap of mean phenotype rank for 20 traits (x-axis) among five types of tobacco named after their role in agronomical practices. **B).** Genomic distribution of 95,308 SNPs in *N. tabacum* genome called from the genotype-by-sequencing approach. **C), D)** Genetic structure of 5196 *N. tabacum* lines from principle component analysis. Samples are colored according to the types given after agronomical practices. **E).** Distribution of pairwise IBS values in different tobacco types. Fixation index (F_{st}) between different tobacco types **F).** and groups defined by genetic structure **G).**

A comprehensive catalogue of the allelic variations underlying variation of plant morphological, physiological, metabolic and disease-resistance traits

With the availability of genome-wide markers and historical phenotypic records on over 5000 lines, we used genome-wide association analysis (GWAS) to perform a comprehensive catalogue of the allelic variation underlying variation in 43 traits (Fig 2A/Fig 3A) with intermediate- to high-heritability (h^2 = 0.15-0.82, Supplementary Table 8-9). 178 significant maker-trait associations were identified for 39 traits (Fig 3B, Supplementary Fig S5), and several highly potential associations were detected and explored further in the following sections. P values and additive effects for the QTLs were summarised in Supplementary Table 10. Although the linkage equilibrium is too extensive to directly pinpoint the candidate genes underlying each association peak, we found a few genes previously identified as QTLs[12,14–18] or annotated as potential candidates nearby eight peaks (Dashed lines in Fig 3B). In addition, we conducted a review of the literature [12,14–18] to determine the relative position of previously reported quantitative trait loci (QTL; Supplementary Table 11) at the ZY300 genome assembly and discovered that 95% of the detected QTL in our study were novel. Altogether, these marker-trait associations constitute a comprehensive genotype-to-phenotype map (Fig 3D, Supplementary Table 12), providing a roadmap for future genetic studies in this model plant species.

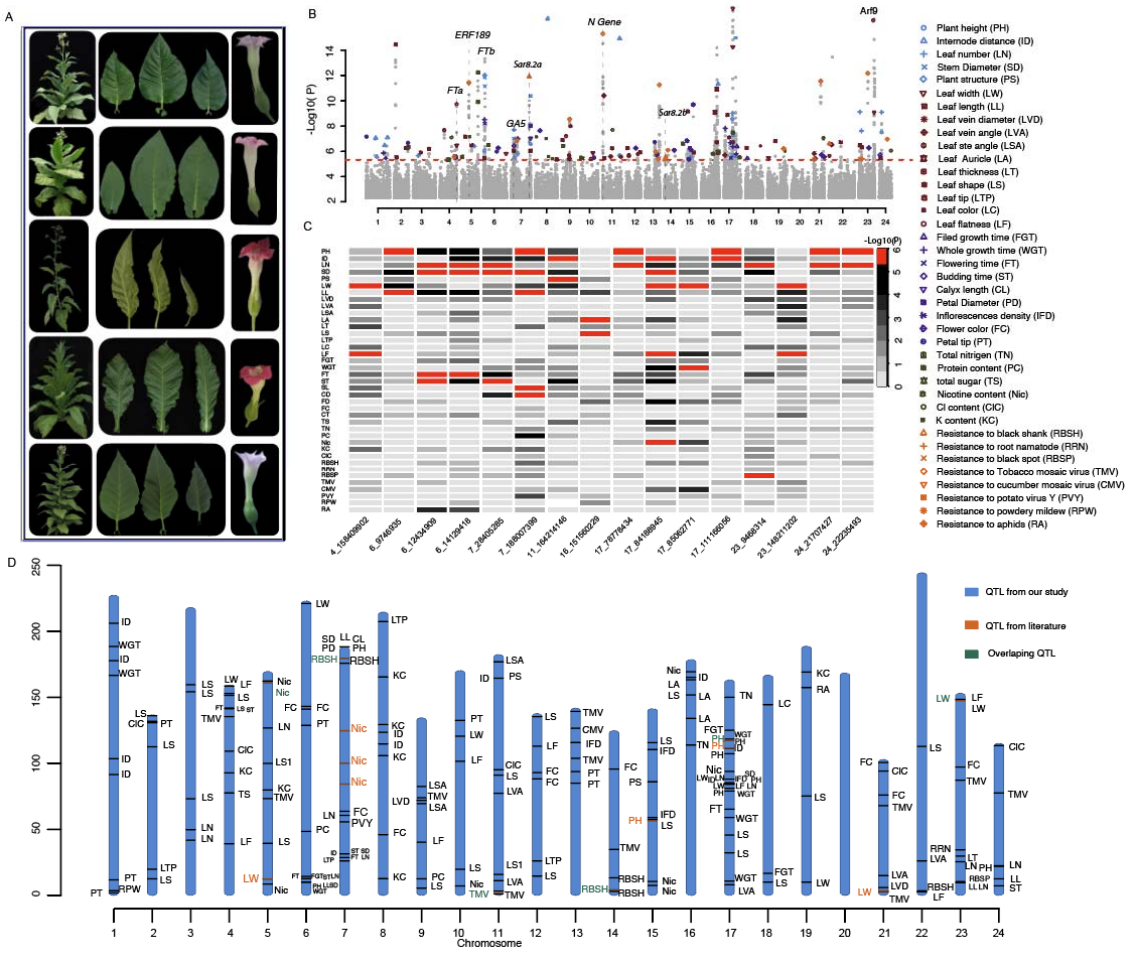

Fig 3. Association results for 43 plant morphological, physiological, metabolic and disease-resistance traits. A). Phenotype diversity of plant architecture, leaf morphology and flowering morphology and colour among five selected plants. The vertical dashed grey lines highlight the genomic position of potential candidate genes. **B).** Manhattan plots of GWAS results from 43 GWAS scans. The red horizontal dashed lines indicate the Bonferroni-corrected genome-wide significance threshold. The vertical dashed grey lines highlight the genomic position of 16 SNPs associated with more than two traits. **C).** A heatmap illustrating the p values of 16 SNPs detected for more than two traits. Each cell represents the -Log10 (p values) 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), previous reports (skyblue), and both (dark green). Previously reported QTLs spanning more than 10Mb and QTLs without probe sequences were not included here.

We detected 32 QTLs (Supplementary Table S10/S12) for four plant architecture traits, plant height (PH), internode distance (ID), leaf number (LN) and stem diameter (SD) that were moderately correlated (Fig 3B, Supplementary Fig S6). Overall, taller plants tend to have more leaves, wider stems and larger internode distance (Fig 4A, Supplementary Fig S4) indicating a shared molecular basis underlying plant architecture development. In consistence with this, four QTLs were associated the variation of two or more plant architecture traits (Fig 3B/3C). For example, the QTL (Chr7: 28 405 285 bp) located at the proximal end of chromosome 7 was associated with variation of PH, LN and SD (Fig 4B). Allele G at this locus showed a dominance effect over allele A (Fig 4D), whereby GG genotype at this locus increased ID, LN and SD by 0.22 ± 0.04 cm ($P= 6.32 \times 10^{-6}$), 0.79 ± 0.15 ($P= 4.02 \times 10^{-8}$) and 0.29 ± 0.05 cm ($P= 8.60 \times 10^{-11}$), respectively (Fig 4D/E/F). Although the linkage disequilibrium (LD) at this region extends for several mega bases, we discovered a potential candidate, NtZY07G00865, a homology to *Arabidopsis thaliana* gibberellin 20-oxidase (AT4G25420, GA5) regulating plant developmental process[19], located inside the association peaks.

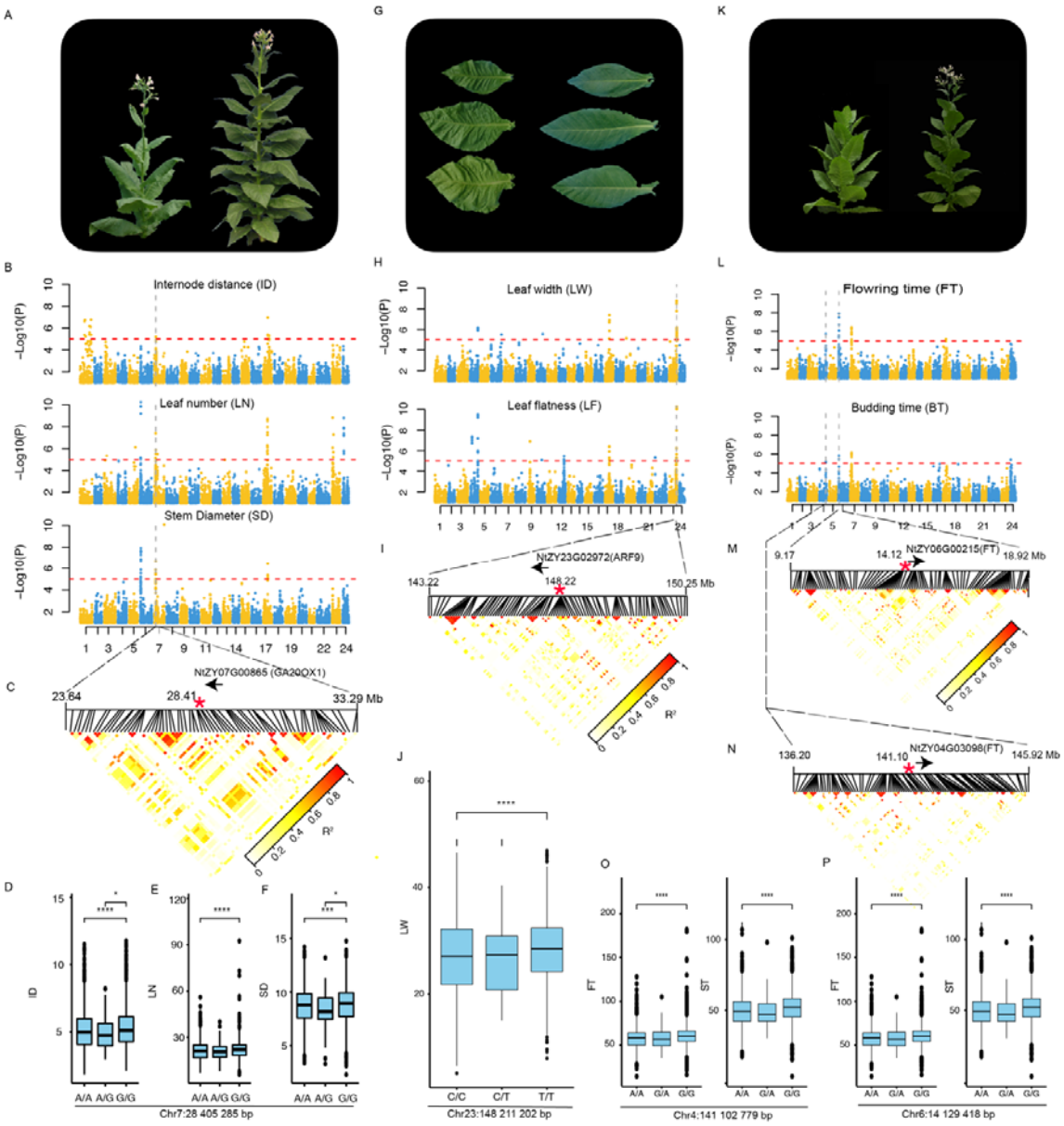

Fig 4. GWAS identify potential candidate genes for plant architecture, leaf morphology and flowering time variation. Graphical illustration of the variation in plant architecture **A),** leaf morphology **G),** and flowering time **K)** traits among *Nicotiana tobacco*. Association results for internode distance, stem diameter and leaf number **B)**, leaf width, leaf flatness **H),** flowering time and bolting time **L). C).** LD heatmap and literature review pinpointed NtZY07G00865, a homology to *Arabidopsis thaliana* AT4G25420 (*GA5*) involved in the later steps of the gibberellin biosynthetic pathway as a potential candidate underlying the association peak at chromosome 7: 28 405 285 bp. **I).** LD heatmap and literature review pinpointed NtZY23G02972, a homology to *Arabidopsis thaliana*, auxin response factor 9

(*Arf9*), regulation cell division activity, as a potential candidate underlying the association peak at chromosome 23: 148 211 202 bp. **M), N),** LD heatmap and literature review pinpointed NtZY04G03098 and NtZY06G00215, both are homology to *Arabidopsis thaliana* flowing time locus T involved in flowering time regulation, as a potential candidate underlying the association peak at chromosome Chr6:14 129 418 bp, Chr4:141 102 779 bp, respectively. **D), E), F), J), O), P)** Genotype-to-phenotype map for the QTLs associated with the corresponding traits.

Strong agronomical interests in leaf usage resulted in an excess amount of diversity in leaf morphology (Fig 3A, Fig 4G) traits. Here, we detected 65 QTLs for 12 leaf morphology traits (Supplementary Table S8), including leaf width (LW), leaf length (LL), vein diameter (VD), leaf stem angle (LSA), leaf vein angle (LVA), leaf auricle (LA), leaf thickness (LT), leaf shape (LS), leaf tip (LTP), leaf color (LC), leaf serration (LSR), and leaf flatness (LF). Overall, the phenotypic correlation among these traits was relatively low (median Spearman correlation = 0.05, Fig S3). However, 3 QTLs were simultaneously associated with LW and LF (Fig 3B/Fig 4H). One of these QTL (Chr23: 148 211 202 bp) is located at the end of chromosome 23. Allele T at this locus increased the LW by 1.11 ± 0.18 ($P= 1.66 \times 10^{-9}$, Fig 4J) and increased the LF by 0.2 ± 0.02 units ($P = 1.58 \times 10^{-18}$). Similar to the peak described in the previous section, LD at this region was too extensive to directly pinpoint candidates, however, we found NtZY23G02972, a homology to *Arabidopsis thaliana* auxin response factor 9 (*Arf9*), regulating cell division activity, as a potential candidate for further validation (Fig 4I).

Regulation of flowering time variation attracted broad interest due to its importance in plant development, ecological adaptation and breeding application. Here, four QTLs were detected for flowering time (FT) and budding time (BT) variation (Fig 4 K-L). Among these, two QTLs located at the end of chromosome 4 (Chr4:141 102 779 bp) and proximal end of chromosome 6 (Chr6:14 129 418 bp) attracted our attention (Fig 4L). Both QTLs included a candidate gene homologous to the *Arabidopsis thaliana* flowering time locus T (*FT*) and these two chromosome regions were synteny blocks from the S and T subgenomes (Fig 4 M/N, Fig 1B). This suggests that the function of both homology genes was retained after the merging of two subgenomes.

In addition to the association peaks described above, we detected additional QTLs associated with leaf metabolites and resistance to several virus-/bacterial-/fungal-induced diseases (Supplementary Table S8). A few of them were tagging genes with known functions reported in previous studies[3,20]. For example, the QTL located on the proximal end of chromosome 11 (Chr11:2 116 418 bp) associated with TMV resistance was tagging *N* gene, a TIR-NBS-LRR protein introgressed through interspecific hybridization with *N. glutinosa*[3,20]. In addition, two QTLs, located on chromosome 7: 175, 557, 338bp and chromosome 14: 2,809,268bp , associated with resistance to black shank (RBSH) was tagging *Sar8.2a* and *Sar8.2b* (Fig 3B) that were detected and fine-mapped previously[17]. Overall, these marker-trait associations represent a comprehensive catalogue of the allelic variations underlying variation of plant morphological, physiological, metabolic and disease-resistance traits, providing a deeper understanding of complex regulation and several candidates for future genetics research in this model plant species.

Fine-mapping the causal variants underlying LW variation and functional characterization of the candidate gene Arf9

Motivated by broad interests in plant leaf development and its importance in light capture efficiency, we attempted to fine-map the QTL at chromosome 23, 148 211 202 bp associated with LW variation. A population of near isogenic lines (NIL) was created by crossing two parents segregating at this QTL, Samsum (P1, T/T at Chr23: 148 211 202 bp) with wide leaves and K326 (P2, C/C at Chr23: 148 211 202 bp) with narrow leaves. Four generations of backcrossing were performed with K326 as the recurrent parent to create a BC4 population and another four generations of self-pollination were performed to generate BC4F4 individuals. Based on the pattern of recombination and phenotype measurements for recombinant genotypes from BC4F2 and BC4F4 (Fig 5A, Supplementary Table S13/S14), the QTL was fine-mapped to a 134kb region between two markers, Chr23: 147231844 bp and Chr23: 147366149 (Fig 5A). There were only two single nucleotide polymorphisms (SNP) between the two parents (Fig 5B) and one gene NtZY23G02972 at this region. NtZY23G02972 was a homology of *Arabidopsis thaliana* auxin response factor 9 (*Arf9*), which widely exists in a number of crops, but has not been functionally characterized. CRISPR-Cas9 knock lines of NtZY23G02972 in *Nicotiana tabacum var XHJ* showed increased LW by 6 cm (2.74×10^{-3}; Fig 6 C-E, Supplementary Fig S7), suggesting that the non-functional allele of NtZY23G02972 was associated with wider leaf. Previously, structure of

ARF5 was characterized and ARF DNA-binding domains form a homodimer to generate cooperative DNA binding, which is critical for *in vivo* ARF5 function[21]. Here, we found that the second SNP at chromosome 23: 147311281 bp located at the eighth exon of NtZY23G02972 was altering the translated amino acid sequence from Alanine (Ala_{203}) to Prolin (Pro_{203}), and the first SNP was located at the first exon of *Arf9* altering the amino acid sequence from Glutamic acid (Glu_{29}) to Glycine (Gly_{29}). This second SNP was located inside the functional domain forming a homodimer (Fig 5F), while the first SNP was at the proximal end of the protein that is less likely to have an impact on the protein activity. Altogether, these results suggested that a causal variant at chromosome 23, 147311281 bp affected the formation of homodimer by changing the amino acid sequence from Alanine (Ala_{203}) to Prolin (Pro_{203}), and increased leaf width. Our results revealed a novel gene regulating plant leaf development and provided a potential candidate for future functional genetic research.

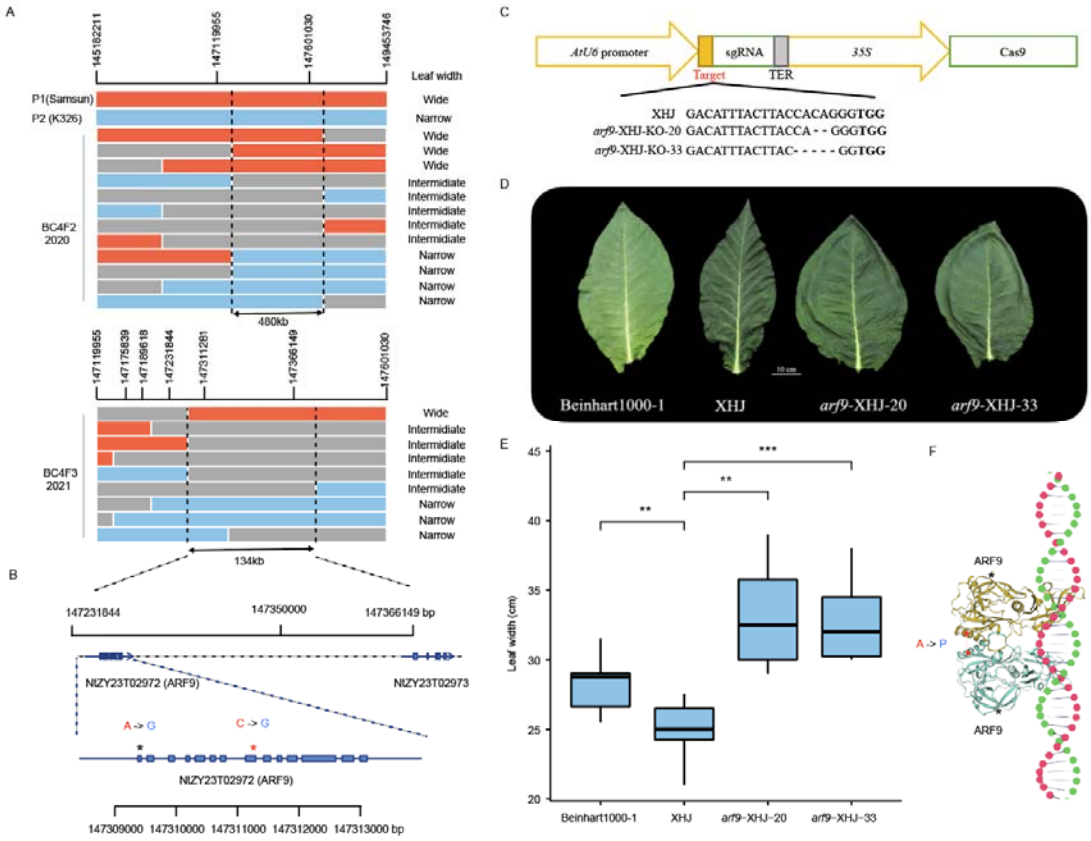

Fig 5. Fine-mapping and functional characterization of the QTL located at chromosome 23, 148 211 202 bp associated with leaf width variation. A). Fine-mapping using near isogenic lines (NIL). Each bar represents a unique genotype at this QTL region. Tomato represents genotype of the donor parent (P1; Samsum) with wider leaves, while skyblue

highlights the genotype of the recurrent parent (P2; K326) with narrower leaf. Unique offspring genotype classes from two generations, BC4F2 and BC4F3, were illustrated using colored bar, and the line of origin were highlighted using tomato, skyblue, grey for homozygous P1, homozygous P2, and heterozygous, respectively. **B).** Gene structure at the fine-mapped region and position of the two SNPs segregating in this population. Causal SNP was highlighted using red star. **C).** Targeted sequences of NtZY23T02972 (*Arf9*) in CRISPR-Cas9 knock lines. **D).** Illustration of the leaf width for a control line (Benhart 1000-1) and targeted line before (XHJ) and after (*arf9*-XHJ-20, *arf9*-XHJ-33) CRISPR-Cas9 knock. **E).** Boxplot of the leaf width for a control line (Benhart 1000-1) and targeted line before (XHJ) and after (*arf9*-XHJ-20, *arf9*-XHJ-33) CRISPR-Cas9 knock. **F).** A SNP altered the protein sequence of ARF from P > M, potentially affecting the formation of a homodimer in DNA binding.

Signature of positive selection and polygenic selection during the process of breeding

When the germplasms were classified into landraces, introduced materials and varieties developed over a period of several hundred years, seven traits, LW, LL, FT, SL, RS, TS and RBSH, were found to be continuously increased/decreased during the process of breeding (Fig 6 A-H). To test for evidence of selection, we evaluated whether the frequencies of favorable alleles associated with each trait increased over time. In total, 46.42% of the QTLs (highlighted using black/red star in Fig 6I) showed increased allele frequency during the process of breeding. However, none of the alleles reached fixation (frequency >0.95), and only 17.00 % of the QTLs showed allele frequencies greater than 0.8, indicating a great potential for future genetic improvement. Among these, 3 QTLs (10.71%) displayed continuously increased allele frequencies from landrace to introduced materials and then to variety (highlighted using red star in Fig 6I). For example, one QTL located on chromosome 23: 148, 211, 202 bp is associated with LW variation. Allele T at this locus increased leaf width by 1.11 cm (Fig 6J; P value = 1.67×10^{-9}), and the frequency of T allele was continuously increased from 0.15 (Landraces) to 0.38 (Introduced lines) and then to 0.47 (Varieties; Fig 6J), suggesting this allele is under positive selection.

In addition, polygenic scores were calculated using alleles with -log10(p-value) greater than 2.5, representing aggregated effects from minor effect alleles. Overall, these minor effect alleles accounted for 23.12% to 95.56% (median = 33.40%, Fig 6K) of the kinship heritability,

suggesting an important contribution from polygenic background to variation of these traits. Taking RS as an example, albeit no significant association was detected at genome-wide significance threshold, aggregated effects from 229 alleles with -log10(p-value) above 2.5 (Fig 6L) explained 43% of the phenotypic variation and 95.56% of the kinship heritability (Fig 6L/M). In total, 37.55% (86) of the alleles showed increased allele frequency from the landrace to introduced materials or variety (Fig 6N). Although the magnitude of frequency increase was small (mean = 0.07, Fig 6N), together they increased the level of reducing sugar from 8.07% to 14.71% from landrace to variety (Fig 6O), demonstrating the power of polygenic adaptation in response to artificial selection. Overall, these results demonstrated the power of leveraging GenBank genomics for dissecting the genetic basis of complex traits variation and evolution during the process of selective breeding and provided a blueprint and guideline for future crop improvement.

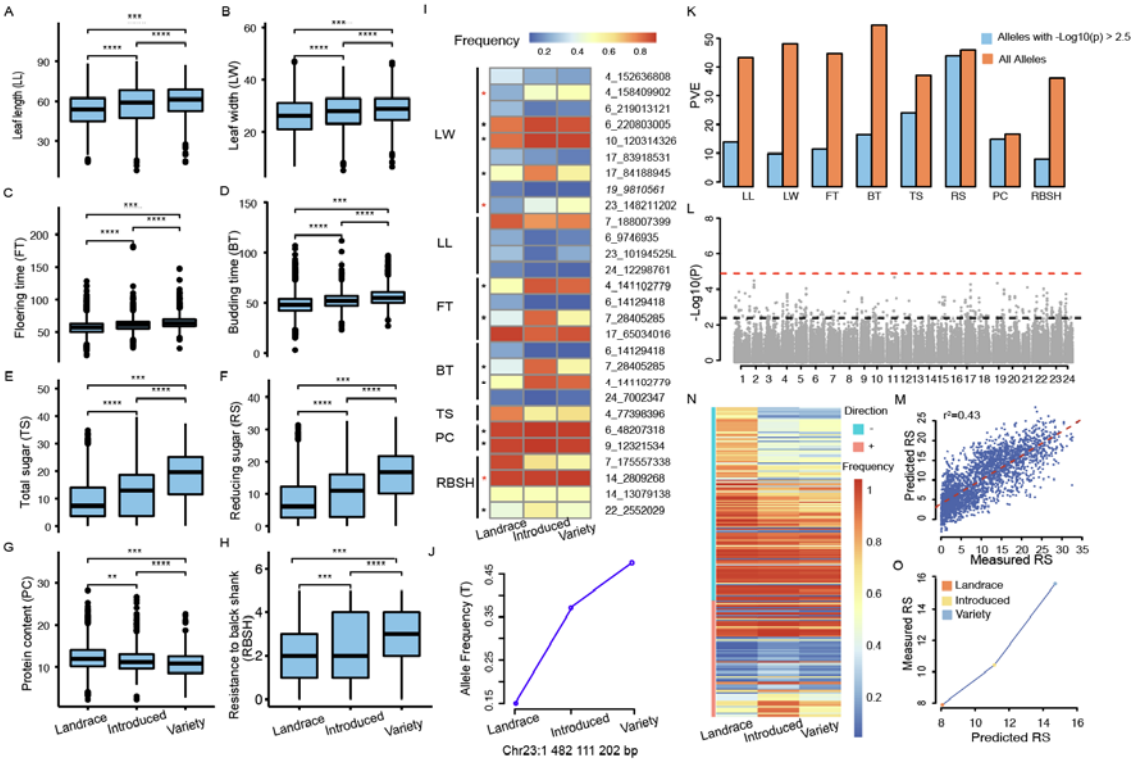

Fig 6. Signature of selection during the process of selective breeding. A-H) Phenotype distribution of eight traits among different categories of materials, landraces, introduced lines, and varieties developed during the process of breeding. **I).** Frequency shifts of 28 QTLs associated with the variation of eight traits. **J).** Frequency shifts of the QTL associated with leaf width. **K).** The proportion of variance (PVE) explained by the alleles with -log10(P) above 2.5 and all the alleles, estimated in a linear mixed model. **L).** Manhattan plot for RS association analysis. **M**). Scatterplot between predicted RS using 229 alleles with -log10(P)

above 2.5 and measured alleles. **N).** Frequency shifts of the 229 alleles with -log10(P) above 2.5 in genome-wide association analysis of RS. **O).** Relationship between the mean of predicted RS level from 229 alleles with -log10(P) above 2.5 and measured RS for three categories of plant materials developed during the process of breeding.

Discussion

In contrast to *Nicotiana benthamiana*, an ancient allotetraploid species (~ 6 Mya) in the *Nicotiana* genus that experienced significant reduction in chromosome number (24 >19) and genome size (~50%)[22] after polyploidisation. Over 90% of the parental genomic sequences is retained in the *Nicotiana tabacum* genome and only 12 chromosome rearrangement events were identified. This suggests that genome downsizing occurs on an evolutionary scale of several million years, and stabilizing chromosome pairing during meiosis by chromosome rearrangement were likely to be the primary task in such a young polyploid (0.2Mya). The high sequence similarity and high expression divergence between 1276 homologous genes (Supplementary Fig S3) indicate that there was likely significant epigenetic remodeling for subgenome coordination.

In curing crops, such as tobacco and coffee, a primary target of breeding is balancing quality with certain morphological and phenology traits, whereby directional selection is relatively rare in comparison to that in field or horticulture crops. In addition, the history of selective breeding in tobacco is considerably short, being only ~ 200 years for flue-cured varieties and ~ 500 years for all other varieties. These likely explain the observed low divergence between major tobacco types, which in turn facilitated high statistical power for GWAS in *Nicotiana tabacum*. Future research with resequencing data is required to investigate the signature of selection in a finer resolution.

We provided a comprehensive genotype-to-phenotype map in this model plant species, and demonstrated the power of this roadmap in plant functional genetics by fine-mapping one of the QTLs. For example, leaf width is an important morphological trait associated with photosynthesis and is of great importance in light capture efficiency. By leveraging the genotype-to-phenotype map, we detected, fine-mapped and functionally validated a novel gene, ARF9, regulating leaf width. Homologous of *Arf9* is widely present in a number of field crops, such as maize, rice, wheat etc., that has 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 the first chromosome-level assembly of the *Nicotiana tabacum* allotetraploid genome in the *Nicotiana* genus, and the global genetic and phenotypic polymorphisms by sequencing and phenotyping an entire GenBank of 5196 *Nicotiana tabacum* germplasm resources. The extensive genotype and phenotype datasets, marker-trait associations, and candidate genes presented in this study will serve as a community resource for accelerating future comparative genomics, plant functional genomic and crop improvement research.

Materials and Method

Plant materials, DNA extraction and sequencing

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

Low-quality paired reads (reads with ≥ 10% unidentified nucleotides (N); > 10 nt aligned to the adapter, > 20% bases having phred quality <5 and putative PCR duplicates generated in 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.

Genome assembly and scaffolding

For initial genome assembly, PacBio long sequence data were self-corrected to generate pre-assembled reads, and assembled by the Overlap-Layout-consensus algorithm [23]. The assembled contigs were further polished with Illumina short-read sequence data by Pilon pipeline[24]. The polished contigs were further connected to generate super scaffolds using 10 x Genomics linked-reads data using fragScaff software[25]. Scaffolds were anchored by linkage information, restriction enzyme site, and string graph formulation with the package LACHESIS[26]. Hi-C data were used to further improve the quality of assemblies with HiC-Pro software (v.2.10.0). The placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were adjusted manually. A final *de novo* assembly was generated with Bionano Solve (v3.5.1). Chromosomes were named by mapping SSR marks from a widely used linkage map[27] to unify the physical map and linkage map.

Repeats and gene annotation

Extensive de-novo TE Annotator (EDTA v1.9.3)[28], a pipeline combining homology-based and de novo prediction methods using LTRharvest[29], LTR FINDER[29], LTR retriever[30], TIR-learner[31], Helitron Scanner[32], and Repeat Modeler[33]) were used to annotate the TE, 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 *Nicotiana 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), *Ghirsutum hirsutum* (GCF_000987745.1) *Arabidopsis thaliana* (GCA_000001735.1) and *Vitis vinifera* (http://jul2018-plants.ensembl.org/Vitis_vinifera/Info/Index) were used as queries to search against ZY300

genome using GeneWise (v.2.4.1)[34]. For transcriptome-based gene prediction, trimmed RNA-sequencing reads from stem, root, leaf, anther, flower, and axillary bud were *de novo* assembled using Trinity (v2.1.1) [35]. To further improve the annotation, RNA-Seq reads from different tissues were aligned to ZY300 genome using TopHat (v2.0.11) with default parameters to identify exons and splice junctions. The alignment was used as input for Cufflinks(v2.2.1) [36]for transcript assembly with default parameters. For ab initio prediction, we used Augustus[37], GlimmerHMM[38], SNAP[39], GeneID (v1.4)[40] and Genscan[41] to predict gene structure. Finally, all predictions of gene models generated from these approaches were integrated into a consensus gene set using EvidenceModeler (v1.1.170)[42]. After prediction, PASA [43] 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[44]. Motifs and domains were annotated using InterProScan (release 5.32-71.0)[45]. Gene ontology (GO) terms and KEGG pathways of predicted sequences were assigned by InterProScan and KEGG Automatic Annotation Server [45,46], respectively.

Subgenome partitioning and synteny analysis

Illumina short-read sequences from *N. sylvestris* (Accession ID: ERR274529), *N. tomentosiformis* (Accession ID: ERR274543) were downloaded from NCBI SRA database and aligned to the ZY300 reference genome using bwa-mem[47]. Samtools[48] was used to count the number of reads mapped to ZY300 genome from each of the two parental genomes, and average read depth in one Mb window was calculated using a customised R script. First, a preliminary parental origin assignment was made using the window-averaged read ratio. Second, for windows containing breakpoints, manual curation was performed in Integrative Genomics Viewer[49] (IGV) to pinpoint the exact location. WGDI[50] was used to detect syntenic blocks based on collinear genes, and R package circlize[51] was used to illustrate blocks between subgenomes after merging large syntenic blocks.

Subgenomes expression bias analysis

400 RNA sequencing datasets were randomly selected from NCBI SRA database (Accession ID available in Supplementary Table S16) to quantify the level of expression bias for

homology genes from two parental genomes. Raw data were downloaded and trimmed using fastp[52] following the default parameter, and subsequently aligned to ZY300 reference genome using HISAT2[53]. The expression level was quantified as FPKM using Stringtie[54]. Fold change between homologous genes was calculated as the ratio between averaged FPKM values across the 400 samples.

Genotype by sequencing, read mapping, and variant calling

DNA extraction was performed using CTAB methods. 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 at the Illumina NovaSeq platform. Raw reads were trimmed using fastp[52] following the default parameter, and subsequently aligned to ZY300 reference genome using bwa[47]. Samtools[48] mpileup was used for variant calling. The called SNPs were initially filtered for read depth, call rate, and minor allele frequency with vcftools[55] following averaged read DP > 2 and MAF > 0.03, individual missing rate < 0.5, site missing rate <0.5. The filtered genotypes were then imputed using beagle[56]. The imputed SNPs were further filtered for a minor allele frequency greater than 0.03 and 95,308 SNPs remained for downstream analysis.

Population structure analysis

To further assess the relatedness between individuals, principle component analysis was performed using the identify-by-state (IBS) genetic distance matrix calculated in PLINK (v1.90)[57] and principal component analysis was performed using *princomp* function in base R. After grouping samples, VCFtools (v0.1.16)[55] was used to calculate the population differentiation coefficient (F_{st}) among major tobacco types and genetic clusters.

Genome-wide association analysis

Genome-wide analysis (GWA) was conducted for 43 traits (phenotype measurements were available in Supplementary Table S17) using a linear mixed model implemented in the *mlma* module from GCTA[58]. A subsequent conditional analysis implemented in the *cojo* module from GCTA[58] was performed to screen for independent association signals. As the linkage disequilibrium (LD) was extensive in this population, assuming all tested markers were statistically independent and deriving a Bonferroni-corrected significance threshold is too conservative. We, therefore, estimated the effective number of independent markers (Me)[59]

and derived a less conservative threshold following 0.05/Me (1. x 10^{-5} equivalent to -Log_{10} (*p*-value) = 5).

Polygenic selection analysis and genomic prediction

We predicted the aggregated effects from all the SNP with a -Log10 (p value) above 2.5 using the following model.

Y= Xβ + Zu + e (1)

Y is a column vector of length n containing phenotype measurements. X is a matrix of n rows and 1 column on 1s, representing the population mean. u is a random effect vector of the polygenic effects (polygenic scores) representing the aggregated effects from all the SNPs for the n individuals. Z is the corresponding design matrix obtained from a Cholesky decomposition of the kinship matrix G, estimated with the markers with -Log10 (p value) above 2.5 using the *A.mat* function in R package rrBLUP [60]. The Z matrix satisfies ZZ'=G, therefore, u ~ N (0, $I\sigma_g^2$). e is the residual variance with e ~ N (0, $I\sigma_e^2$). The proportion of variance explained by these SNPs was estimated as $\sigma_g^2/(\sigma_g^2+\sigma_e^2)$. Model fit was performed in the R package rrBLUP [60].

Fine-mapping and experimental validation of candidate gene

A population of near isogenic lines (NIL) was developed by marker-assisted backcrossing using Samsun (P1, T/T at Chr23: 148 211 202 bp) as the wide leaf donor parent and K326 (P2, C/C at Chr23: 148 211 202 bp) as the narrow leaf recurrent parent. First, a total of 1694 BC_4F_2 individuals from the K326×Samsun population were genotyped using five markers from Chr23:145182211bp and Chr23:149453746bp (Supplementary Table 13) and then selected recombinants were self-pollinated to create BC_4F_3 lines for further evaluation. This narrowed the QTL to a region between Chr23:147119955 and Chr23:147360000bp. An additional 3017 BC_4F_3 individuals derived from the population were further screened using seven SNP markers in this narrow region (Supplementary Table 13) and the BC_4F_4 lines of recombinants were phenotyped for further fine mapping. Two sgRNAs (small guide RNAs, Fig 5C) were designed to target the candidate gene, NtZY23T02972, based on the assembled ZY300 genome. The vectors were constructed and transformed into the receptor XHJ. The genotype of gene-edited lines was identified by PCR amplification and Sanger sequencing. For fine mapping, LW were recorded by measuring at least 15 plants per line during flowering time at Sanya (E 108.56, N 18.09), China. For transgenic experiments, phenotypes

of knockout lines and the wild type were investigated at Sanya (E 108.56, N 18.09), China. Each experiment had two replicates with at least five independent plants per replicate. The mean LW of at least five plants per replicate were used for further analyses.

Funding

This research was funded by the National Science Foundation of China (32200503), the Crop Germplasm Resources Protection, Utilization Special Grant from the Ministry of Agriculture and Rural Affairs of the People's Republic of China, and National Science and Technology Resource Sharing Service Platform (NCGRC-2023-18).

Author Contributions

A.Y, L.C and Y.C designed and supervised this study. Y.Z, S.C, M.R, G.L, Y.L collected the data and implemented the analysis. H.S contributed in the statistical analysis. Y.L, X.Z, Y.T, C.J, Z.L, D.L, Y.S, Z.X, R.G conducted the field experiments and phenotype measurements. Q.F, Y.W collected the germplasms. Y.Z, S.C and L.C wrote the manuscript.

All authors have read and approved the manuscript, and have no conflicts of interest to declare.

Acknowledgments

We thank Yan Ji for her efforts in improving the Figures and Tables. We thank Wei Zhao and Ying Guo for their constructive comments to improve the manuscript. We thank Shilin Tian for his valuable input during the design and implementation of this study.

Data availability

All sequence data will be available at NCBI upon acceptance of this manuscript.

All phenotype measurements and meta data of samples are included in Supplementary Tables.

Reference

1. Eriksson, T. & Jonasson, K. Nuclear division in isolated protoplasts from cells of higher plants grown in vitro. *Planta* **89**, 85–89 (1969).
2. Clausen, R. E. & Mann, M. C. Inheritance in Nicotiana Tabacum: V. The Occurrence of Haploid Plants in Interspecific Progenies. *Proc. Natl. Acad. Sci.* 121–124 (1924).
3. Holmes, F. O. Inheritance in tobacco of an improved resistance to infection by tobacco mosaic virus. *Virology* **12**, 59–67 (1960).
4. Venkateswarlu, K. & Nazar, R. N. Evidence for T–DNA Mediated Gene Targeting to

Tobacco Chloroplasts. *Bio/Technology* **9**, 1103–1105 (1991).

5. Angell, S. M. Consistent gene silencing in transgenic plants expressing a replicating potato virus X RNA. *EMBO J.* **16**, 3675–3684 (1997).

6. Gunasekaran, B. & Gothandam, K. M. A review on edible vaccines and their prospects. *Brazilian J. Med. Biol. Res. = Rev. Bras. Pesqui. medicas e Biol.* **53**, e8749 (2020).

7. Yao, J., Weng, Y., Dickey, A. & Wang, K. Plants as Factories for Human Pharmaceuticals: Applications and Challenges. *Int. J. Mol. Sci.* **16**, 28549–28565 (2015).

8. Goodin, M. M., Zaitlin, D., Naidu, R. A. & Lommel, S. A. Nicotiana benthamiana□: Its History and Future as a Model for Plant–Pathogen Interactions. *Mol. Plant-Microbe Interact.* **21**, 1015–1026 (2008).

9. Sierro, N. *et al.* The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* **5**, 3833 (2014).

10. Sierro, N. *et al.* Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. *Genome Biol.* **14**, R60 (2013).

11. 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).

12. 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).

13. 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).

14. 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).

15. 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).

16. 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).

17. 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).

18. 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).

19. Zheng, F., Wang, Y., Gu, D. & Liu, X. Histone Deacetylase HDA15 Restrains PHYB-Dependent Seed Germination via Directly Repressing GA20ox1/2 Gene Expression. *Cells* **11**, 3788 (2022).

20. Marathe, R., Anandalakshmi, R., Liu, Y. & Dinesh-Kumar, S. P. The tobacco mosaic virus resistance gene, N. *Mol. Plant Pathol.* **3**, 167–172 (2002).

21. Boer, D. R. *et al.* Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. *Cell* **156**, 577–89 (2014).

22. Bombarely, A. *et al.* A draft genome sequence of Nicotiana benthamiana to enhance molecular plant-microbe biology research. *Mol. Plant. Microbe. Interact.* **25**, 1523–1530 (2012).

23. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).

24. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

25. Adey, A. *et al.* In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).

26. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).

27. 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).

28. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

29. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).

30. 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).

31. 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).

32. 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.* **111**, 10263–10268 (2014).
33. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
34. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
35. 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).
36. Ghosh, S. & Chan, C.-K. K. Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods Mol. Biol.* **1374**, 339–361 (2016).
37. 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).
38. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–9 (2004).
39. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
40. Guigó, R. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comput. Biol.* **5**, 681–702 (1998).
41. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
42. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
43. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–66 (2003).
44. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
45. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–40 (2014).
46. Aoki-Kinoshita, K. F. & Kanehisa, M. Gene Annotation and Pathway Mapping in KEGG. in *Comparative Genomics* 71–92 (Humana Press, 2007). doi:10.1385/1-59745-515-6:71.

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. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–92 (2013).
50. 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).
51. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
52. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
53. 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).
54. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–5 (2015).
55. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
56. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
57. 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).
58. 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).
59. 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**, 747–56 (2012).
60. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* **4**, 250–255 (2011).