#### **ORIGINAL ARTICLE**



# Quantitative trait locus mapping and genomic selection of tobacco (*Nicotiana tabacum* L.) based on high-density genetic map

Zhijun Tong<sup>1</sup> · Zhihui Xiu<sup>2,3</sup> · Yao Ming<sup>2,3</sup> · Dunhuang Fang<sup>1</sup> · Xuejun Chen<sup>1</sup> · Yafei Hu<sup>2,3</sup> · Juhong Zhou<sup>2,3</sup> · Weiming He<sup>2,3</sup> · Fangchan Jiao<sup>1</sup> · Chi Zhang<sup>2,3</sup> · Shancen Zhao<sup>2,3</sup> · Han Jin<sup>2</sup> · Jianbo Jian<sup>2,3,4</sup> · Bingguang Xiao<sup>1</sup>

Received: 2 February 2021 / Revised: 7 August 2021 / Accepted: 12 September 2021 / Published online: 29 October 2021 © The Author(s) 2021

#### Abstract

Tobacco (*Nicotiana tabacum* L.) is an economic crop and a model organism for studies of plant biology and genetics. As an allotetraploid plant generated from interspecific hybridization, tobacco has a massive genome (4.5 Gb). Recently, a genetic map with 45,081 single nucleotide polymorphism (SNP) markers was constructed using whole-genome sequencing data for a tobacco population including 274 individuals. This provides a basis for quantitative trait locus (QTL) mapping and genomic selection, which have been widely applied to other crops but have not been feasible in tobacco. Based on this high-density genetic map, we identified QTLs associated with important agronomic traits, chemical compounds in dry leaves, and hazardous substances in processed cigarettes. The LOD values for major QTLs were highest for agronomic traits, followed by chemical compounds and hazardous substances. In addition to the identification of molecular markers, we evaluated genomic selection models and found that BayesB had the highest prediction accuracy for the recombinant inbred line population. Our results offer new insights into the genetic mechanism underlying important traits, such as agronomic traits and quality-related chemical compounds in tobacco, and will be able to support the application of molecular breeding to tobacco.

Keywords Tobacco · Molecular breeding · Quantitative trait locus · Agronomic trait

— Fangchan Jiao jfc99002@163.com
Chi Zhang zhangchi2@bgi.com
Shancen Zhao zhaoshancen@bgi.com
Han Jin jinhan@bgi.com
<sup>1</sup> Key Laboratory of Tobacco Biotechnological Breeding, National Tobacco Genetic Engineering Research Center,
Yunnan Academy of Tobacco Agricultural Sciences, Kunming 650021, Yunnan, China
<sup>2</sup> BGI Genomics, BGI-Shenzhen, Shenzhen 518083, Guangdong, China
<sup>3</sup> Center for Plant and Animal Genomics Engineering Researc of Guangdong Province, Shenzhen 518083, Guangdong,
China <sup>4</sup> Department of Biotechnology and Biomedicine. Technical
University of Denmark, 2800 Lyngby, Denmark
_

#### Introduction

Tobacco (Nicotiana tabacum L.) is an important industrial crop and is a model for transient expression. As a commercial crop, tobacco is widely cultivated in more than 120 countries (Sierro et al. 2014). According to the FAO, approximately 6.7 million tons of tobacco are produced worldwide, and China accounts for ~ 39.6% of the total yield (Geneva: World Health Organization 2017). Tobacco goods are mainly derived from dried leaves, which are processed as cigarettes, cigars, chewing tobacco, pipe tobacco, and shisha tobacco. It is predominantly consumed for the addictive and stimulant alkaloid nicotine. Tobacco essentially started the era of plant genetic engineering (Vasil and Hildebrandt 1965; Hoekema et al. 1983). It is a model plant for genetic detection, pathology investigations, and biotechnology. The tobacco leaf has been used as a model for studies of mechanisms underlying diseases in the family Solanaceae and in other crops (Wang and Bennetzen 2015; Wang and Balint-Kurti 2016). It is commonly used to evaluate protein subcellular localization via transient expression in epidermal cells (Liu et al. 2013; Sun et al. 2015; Xiu et al. 2016) and for functinal analyses of genes in other species (Bao et al. 2017; Pagliano et al. 2017; Lee and Kim 2018). Despite its wide utilization as a model plant, the genetic mechanisms underlying important agronomic traits in tobacco are not well-characterized owing to the genome complexity and limited molecular markers (Julio et al. 2006; Lewis et al. 2007). Since the genome assembly of different tobacco cultivars has been reported (Sierro et al. 2014), allotetraploid tobacco (2n = 4x = 48)has been a model system for studies of genomic structure and evolution.

As a member of the family Solanaceae, tobacco is an allotetraploid generated from the interspecific hybridization of Nicotiana sylvestris (2n = 24) and Nicotiana tomentosiformis (2n = 24) (Leitch et al. 2008; Sierro et al. 2013). The genome of tobacco (4.5 Gb) is complex, with numerous repetitive sequences (Kenton et al. 1993). Advances in sequencing technologies and genome assembly methods have enabled the generation of a nearly complete reference genome for allotetraploid tobacco (Thimmegowda et al. 2018; Tong et al. 2020). In addition, the development of molecular genetic markers is increasing (Ganal et al. 2009; Thimmegowda et al. 2018). For example, both restriction site-associated DNA sequencing (RAD-seq) and whole genome resequencing approaches have been employed to identify single-nucleotide polymorphisms (SNPs) and to construct linkage maps (Xiao et al. 2015; Thimmegowda et al. 2018). Recently, a genetic map with 45,081 SNPs markers has been constructed by whole genome sequencing using a tobacco population including 274 accessions (Tong et al. 2020). This high-density genetic map provides a foundation for tobacco quantitative trait locus (QTL) studies.

For the characterization of complex traits, the combination of QTL mapping and molecular marker-assisted selection is an effective strategy. This approach has been applied to many crops, including maize (Liu et al. 2014; Zhang et al. 2017), rice (Dixit et al. 2017; Hu et al. 2018), and wheat (Prat et al. 2017; Tura et al. 2020). In Solanaceae, this approach has been used in tomato (Bennewitz et al. 2018; Diouf et al. 2018). However, the application of QTL mapping to tobacco is limited by the genome complexity and insufficient molecular markers (Julio et al. 2006; Lewis et al. 2007). Recently, QTL mapping has been used to identify disease resistance-associated QTLs. For example, QTLs for resistance to black shank disease (caused by Phytophthora nicotianae), brown spot (caused by Alternaria alternata), soilborne disease, cucumber mosaic virus, and Phytophthora nicotianae have been identified in tobacco via genetic linkage mapping and association analyses (Tong et al. 2012; Drake-Stowe et al. 2017; Zhang et al. 2018; Cheng et al. 2019; Ma et al. 2019). Moreover, molecular markers associated with other important agronomic traits, such as plant height, leaf number, and leaf color, have been identified by QTL mapping (Lewis et al. 2007; Wu et al. 2014; Cheng et al. 2015). However, analyses of economic traits are limited, and related QTLs are insufficient for molecular breeding, especially for quality improvement of tobacco products. In addition, with a large number of genetic markers, genomic selection (GS), which exploits a reference population with genotypic and phenotypic data to predict performance or genomic breeding values for population inferences, is useful for complex trait breeding not only in model crop species but also in non-model species (Bhat et al. 2016). GS methods have been reviewed for crop breeding and is now a common approach (Wang et al. 2018). However, few studies have utilized GS in tobacco.

The quality of tobacco products is mainly determined by leaf chemical compounds, including aromatic substances, nicotine, tar, total nitrogen, starch, total sugar, reductive sugars, and hazardous substances (e.g., benzo[a]pyrene, phenol, and nicotine) (McAdam et al. 2016). Quality improvement in tobacco is achieved by the optimization of the proportions of aromatic substances, nicotine, and tar and reductions in hazardous substances. However, studies of molecular mechanisms underlying related traits are scarce, despite implications for molecular breeding. In this study, we identified QTLs associated with important agronomic traits, chemical compound contents of dry leaves, and hazardous substances in processed cigarettes using a published high-density genetic map with whole genome sequencing (Tong et al. 2020). QTLs for seven agronomic traits, including plant height, leaf number, stem girth, internode length, length of the largest leaf, width of the largest leaf, and yield, were mapped to linkage group 6 (LG6). Multiple QTLs affecting the contents of 23 chemical compounds in dry leaves and hazardous substances in processed cigarette were mapped to different LGs. Along with the identification of QTLs, we estimated that the GS model (BayesB) had the highest accuracy for the RIL population. These QTLs with functional annotations and the accurate GS pilot test provide new insights into genetic mechanisms underlying important agronomic traits and quality-related chemical compounds in tobacco and offer a foundation for molecular breeding.

#### **Materials and methods**

#### Plant material, genetic linkage map, phenotyping, and trait evaluation

The population, including 271 RIL\_F7 individuals (7th generation from parental lines), the parents (Y3, a landrace variety, and K326, a commercial cultivar with a reference genome), and an F1 individual, was planted in Yanhe and Shilin (Yunnan Province, China) in 2018 and 2019 under natural conditions. We have previously published the genetic linkage map (Tong et al. 2020). In brief, a linkage map spanning 3484.78 cM with 45,081 markers was characterized. This high-density genetic linkage map was used for a QTL analysis.

Seven agronomic traits, including natural plant height (npH), natural leaf number (nLN), stem girth (SG), internode length (IL), length of the largest leaf (LL), width of the largest leaf (WL) and yield per plant (Y), were measured at the Yanhe (N: 24.35; E: 102.54) and Shilin (N: 23.46; E: 103.17) stations in two consecutive years (2018 and 2019, except data for yield were collected only in 2018). The four environmental conditions in which tobacco lines were cultivated were E1 (2018 Shilin), E2 (2018 Yanhe), E3 (2019 Shilin), and E4 (2019 Yanhe). Seven agronomic traits were measured 65 days after planting in the field (First green fruit stage). The heritability analysis of six agronomic traits (two points at 2 years) was performed with Best linear unbiased prediction (BLUP) among replicates using the R package "Ime4" (Bates et al. 2007).

Twenty-three chemical compounds (sugar alkali ratio, shim uke value, total plant alkali, total sugar, reducing sugar, total nitrogen, potassium, chlorine, starch, nicotine, nornicotine, the ratio of nornicotine, anabasine, anatabine, ruti, glucose, fructose, sucrose xanthophyl II, beta-carotene, free citric acid, crude fiber and petroleum ether) were collected in dry leaves of plants at the Yanhe and Shilin stations in 2018. In brief, a sample of C3F grade dry leaves were selected to make 100 g leaves powder for chemical compounds determination. Total plant alkali, total nitrogen, chlorine and starch were measured using continuous flow method. Total sugar and reducing sugar were measured subbing water-soluble sugar determination method. Potassium was determined using fame photometry. Nicotine, nornicotine, the ratio of nornicotine, anabasine and anatabine were measured by GC–MS. Ruti, glucose, fructose, sucrose xanthophyl II and beta-carotene were measured by high-performance liquid chromatography (HPLC). Free citric acid was measured by gas chromatographic method. Crude fiber and petroleum ether were determined by gravimetric method (Julio et al. 2006).

The seven hazardous substances, including benzo[a] pyrene, hydrocyanic acid, phenol, carbon monoxide, tar, nicotine and total particle matter, were collected in the mainstream smoke of cigarettes produced using tobacco planted at Yanhe in 2018. In brief, the cigarette was placed in a constant temperature and humidity box with temperature at  $22 \pm 2$  °C and relative humidity of  $60 \pm 5\%$  for 48 h. After selected by the weight and absorption resistance, total particle matter of smoke was collected using Cambridge filter with the SM450 linear smoking machine (suction time 2 s, suction interval 58 s, suction capacity 35 mL). Benzo[a]pyrene and nicotine were measured by GS-MS. Hydrocyanic acid was measured by continuous flow method. Phenol was measured by HPLC. Carbon monoxide was determined by near-infrared reflectance spectrometry (NIRS). Tar was determined by gravimetric method. Details of related traits are provided in Table S1. Average values for six biological replicates were obtained. The frequency distributions and correlation coefficients of all traits were analyzed using the describe() function of the *psych* module and the *cor()* function of the *spearman* module in R, respectively.

#### QTL analysis

The QTL analysis was performed using  $MapQTL \ 6.0$ . Data for all traits in different environments and years were analyzed separately. The genome-wide LOD threshold (p < 0.05) for each trait was calculated using the PERMUTATION test with 1000 permutations (Churchill and Doerge 1994). LOD values for all significant markers associated with candidate traits were obtained by the regression approach and interval mapping. All mapping markers whose LOD values were equal to or greater than the thresholds were retained and used as cofactors in the multiple-QTL models (MQM) in  $MapQTL \ 6.0$ . All mapping information including chromosomal location, magnitude and environmental situations were generated using R. The markers with LODs greater than the LOD threshold were identified as the optimal final markers.

#### **Candidate gene prediction**

All candidates within the confidence interval were aligned against databases, including KEGG (http://www.genome. jp/kegg/) and NR (NCBI Non-redundant protein), using BLASTp with an *E* value cutoff of 1 E–5. InterPro (Mulder and Apweiler 2007) was used to annotate protein motifs and domains, and Gene Ontology (GO) terms were retrieved (Ashburner et al. 2000).

#### **Genomic prediction**

The accuracy of GS is determined by a few factors, including the population size, GS model, marker density, and relative training set and test set sizes. Four model types, including ridge regression BLUP (rrBLUP), which estimates marker effects by ridge regression (Endelman 2011), and Bayesian regression using mixture models (BayesB assumes that each marker has a unique variance and a proportion  $(\pi)$  of markers has large effects, while most markers have no effect), including BayesA, BayesB, and BayesC (Meuwissen et al. 2001; R development core team, 2011), were investigated for 7 agronomic traits, 23 chemical compounds, and 7 harmful components. For marker density, we tested the prediction accuracy with different gradients of SNPs (1K, 2K, 4K, 7K, 11K, 16K, 22K, 29K, 37K, and all SNPs) with 50 cycles for each group. We evaluated population sizes of 50, 100, 150, 200, 250, and 271 (all) with 50 cycles for each group. For the ratio of the training set to the test set population size, we evaluated 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5, 6, and 10) with 50 cycles for each group.

#### Results

## Trait analysis of the recombinant inbred line population

Phenotypic data were obtained from two field stations in 2 years, yielding four data sets for RIL population traits. For seven agronomic traits, 26 environmental trait data sets (excluding yield data for 2019) were collected. For example, the average tobacco natural plant height (npH) ranged from 144.94 cm (E3) to 196.44 cm (E2). The average tobacco natural leaf number (nLH) ranged from 26 (E1) to 29 (E2). The average tobacco stem girth (SG) ranged from 7.96 cm (E4) to 10.16 cm (E2). Detailed information for other traits is shown in Table S2. These seven agronomic traits varied among environments. The Spearman correlation coefficients indicated that the relationships among these traits were generally significant ( $p \le 0.01$ ), indicating a normal distribution among populations (Fig. S1). To evaluate the genetic factors contribution in the trait variation, we performed the heritability analysis among six agronomic traits which had data sets at two stations in 2 years. The heritability of npH, nLN, SG, IL, LL and WL is 77.6%, 69.02%, 66.45%, 0%, 36.07% and 58.13% with BLUP methods. The result showed that most of the phenotype trait have high heritability except IL (Table 1).

For the contents of 23 chemical compounds in the dry leaf, 46 environmental trait data sets (two stations at 2018) were collected and used for correlation analyses (Fig. S2). Summary statistics and correlation coefficients for these environmental traits (means, SD, and CV in the RIL population) are shown in Table S3. The contents of the same chemical compound in different environments were significantly correlated ( $p \le 0.01$ ). Data for seven hazardous substances were only collected in Yanhe 2018 owing to a limited sample size. The contents of hazardous substances per cigarette are provided in Table S4. The data followed a normal distribution and Spearman correlation coefficients were evaluated

Table 1 Heritability analysis of six agronomic traits with two stations in 2 years

Trait	Range				Mean ± SE	Heritability (%)
	Parents			RIL population		
	YK326	YY3	YF1			
npH (cm)	111.44–193.88	221.31-270.56	153.07-223.56	87.33-275.67	177.74±1.70	77.60
nLN (n)	25.22-28	36.33-36.74	27.80-32	18.5-45.00	$27.96 \pm 0.25$	69.02
SG (cm)	7.28-11	9.9-11.22	8.10-10.83	5.00-13.17	$8.85 \pm 0.07$	66.45
IL (cm)	3.70-7.44	4.98-6.33	4.5-5.55	2.64-7.55	$4.66 \pm 0.04$	0.00
LL (cm)	58.35-73.11	61.92-74.06	60.90-71.5	37.40-92.33	$62.74 \pm 0.39$	36.07
WL (cm)	25.45-34.91	29.48-36.67	27.56-36.53	13.50-45.27	$27.74 \pm 0.25$	58.13

(Fig. S3). SC7 (total particle matter) showed strong, positive, and significant correlations with SC2 (hydrocyanic acid), SC3 (pheno), SC4 (carbon monoxide), SC5 (tar), and SC6 (nicotine) (p < 0.01). These results showed that the correlations between chemical compounds content were significantly weaker than those among agronomic traits.

#### **QTL** analysis

To detect QTLs for each of the three kinds of traits by integrating phenotypic and genetic data for the RIL population, an interval mapping method was used to determine logarithm of odds (LOD) scores (Jansen 1993). Different environmental conditions and the same traits were compared. For agronomic traits, most LOD scores were approximately 3.6–3.8, except the LOD score for the length of the largest leaf (LL) in E1 was 2.3. Thirteen regions, including 43 significant QTLs, were identified in different environments (Table S5). In detail, these QTLs were mapped to linkage group (LG) 1, LG5, LG6, LG8, and LG17 with 2, 2, 5, 3, and 1 region, respectively (Fig. 1). There were 3447 SNP markers corresponding to 171 protein-coding genes distributed on 2651 scaffolds in these 13 QTL regions. Gene annotation (171 genes) revealed that some QTLs encode proteins, such as SMAD/FHA domain-containing protein and glycosyl hydrolase family protein, which are related to growth and development. The QTL results showed that nearly all the seven traits were associated with a major region of LG6 with the highest LOD. The main effect QTL regions for natural plant height (nPH), natural leaf number (nLN), and stem girth (SG) were located in LG6 from 101.14 to 158.72 cM

with LOD values exceeding 25 (Fig. 1 and Table S5). Similar regions were also identified for the length of the largest leaf (LL) with an LOD of 12.5 and yield per plant (Y) with an LOD of 7 (Fig. 1 and Table S5). The identification of the same QTL location for these five agronomic traits was consistent with the high correlations between phenotypic values for these traits. Compared with agronomic traits, the chemical compounds in the dry leaf were more complex, involving more factors. In the QTL analysis of 23 chemical compounds in the dry leaf, 17 regions were identified in different environments with LOD scores of 3.2-4.0 LOD (most frequently 3.6, 3.7, and 3.8) (Table S6). The identified QTLs mapped to linkage group (LG) 1, LG3, LG4, LG6, LG8, LG12, LG14, LG15, and LG22 with 2, 1, 1, 4, 2, 2, 4, 1, and 1 region, respectively (Fig. 2). There were 4301 SNP markers (655 bin markers) distributed on 3423 scaffolds in these regions. In total, 195 genes were annotated. A Gene Ontology (GO) analysis of these genes showed enrichment for metabolic processes (biological process) and catalytic activity (molecular function) (Fig. S4). These terms were related to the biosynthesis of chemical compounds. The LOD values for major OTLs for chemical compounds were significantly lower than those for agronomic traits. The main major QTL for eight chemical compounds, including sugar alkali ratio (CC1), shi Muke value (CC2), total plant alkali (CC3), nicotine (CC10), nornicotine (CC11), the ratio of nornicotine (CC12), anabasine (CC13), and anatabine (CC14), were mapped to the same location from 86.40 to 135.30 cM in LG14 (Fig. 2 and Table S6). The highest LOD value was > 16 in CC3, CC10, CC12, and CC13. The other chemical compounds showed relatively low LOD values,



Fig.1 QTL mapping results for seven agronomic traits in the RIL population in four environmental conditions. Different blocks represent linkage groups (LGs). LOD cutoff values are indicated by

the dotted line. nPH natural plant height, nLN natural leaf number, SG stem girth, IL internode length, LL length of the largest leaf, WL width of the largest leaf, Y yield per plant



Fig. 2 QTL mapping results for chemical compound contents in the dry leaf in the RIL population in two environmental conditions. Linkage groups (LGs) are indicated by different blocks. Dotted lines

represent the cutoff values for LODs. The abbreviations for chemical compounds are defined in Table S1

maybe owing to the complexity of these traits. For the seven hazardous substances in cigarette mainstream smoke, six QTL regions were identified with LOD scores of 3.6-3.8 (Table S7). Four were mapped to LG4 and the remaining two were in LG12 and LG14 (Fig. 3). These six regions contained 1933 SNP markers (208 bin markers) distributed on 1460 scaffolds. The 128 annotated genes included genes encoding enzyme activity proteins, such as synthase, oxidase, and hydrolase, which might be involved in the biosynthesis of harmful substances. The major QTL regions of SC3 (pheno), SC4 (carbon monoxide), SC5 (tar), SC6 (nicotine), and SC7 (total particle matter) were located in LG6 from 123.28 to 158.72 cM. The similar locations were consistent with the strong positive correlations among these traits. The LOD values (~4) for hazardous substances were slightly higher than the cutoff LOD, suggesting that hazardous substance contents were complex traits, making the identification of QTLs more difficult compared with agronomic traits and chemical compounds.

#### Genomic selection analysis

GS has not been reported in tobacco to date. We applied GS to the tobacco RIL population. The effects of marker number, population size, and the ratio of the training set to the test set population size on prediction accuracy were tested

for all of the phenotypic traits. As shown in Fig. 4, as the number of markers increased from 1 to 2k, the prediction accuracy for the yield in YH in the RIL population increased slightly (from 0.155 to 0.181), followed by a slight decrease at 7k (0.114) and a subsequent increase in prediction accuracy at 45.9k (0.202). The population size was a significant determinant of prediction accuracy. When the population size increased to 250, the prediction accuracy nearly reached its maximum value. The optimal ratio for the test set to the training set with respect to prediction accuracy was 10. Four statistical models (rrBLUP, BayesA, BayesB, and BayesC) were evaluated (Fig. 4), with average accuracies of 0.242, 0.262, 0.264, and 0.206, respectively. Significant differences could be identified for the RIL population with average accuracies in rrBLUP and Bayes models. The BayesB model showed the highest accuracy for the RIL population. A similar trend was observed for other phenotype traits, with slight differences.

### Discussion

Using a previously established high-density genetic linkage map (Tong et al. 2020), we identified QTLs associated with yield-related traits, including plant height, leaf number, stem girth, internode length, length of the largest leaf, width of



**Fig.3** QTL mapping results for harmful ingredients in cigarette smoke in the RIL population. Different blocks represent linkage groups (LGs). LOD cutoff values are indicated by the dotted line. The abbreviations for harmful ingredients in cigarette smoke are defined in Table S1





**Fig. 4** Prediction accuracy (total value = 1) of four different statistical models for the yield of YH in RIL populations. **a** Marker number was the variable factor. **b** Population size was the variable factor. **c** 

Ratio of training set to test set population sizes was the variable factor. **d** Different statistical models were used for prediction. Prediction accuracy is calculated as 1 in the Y axis

the largest leaf, and yield of single plants. Compared with previous reports focused on plant height, leaf number, and leaf color (Lewis et al. 2007; Wu et al. 2014; Cheng et al. 2015), we performed a comprehensive analysis of important agronomic traits. In addition, we mapped QTLs for the contents of chemical compounds, providing the first analysis of the genetic control of these quality-related traits in tobacco. We specifically isolated seven harmful ingredients for QTL analyses. Compared with traditional technologies used to reduce harmful ingredients in cigarette smoke, strategies targeting the molecular mechanisms underlying the production of hazardous substances would be an innovation approach to improve safety. The functional annotation of genes corresponding to these QTLs provides new insight into the genetic mechanism underlying important agronomic traits. In addition, these results provide foundational information for both yield and quality improvement in tobacco by molecular breeding.

Among the three trait types, QTLs for agronomic traits showed the highest LOD values and fewer genomic regions. The LODs for hazardous substances in cigarette mainstream smoke were lower than those for chemical compounds in the dry leaf. It is possible that more complex traits showed lower LODs and mapped to more QTLs. Alternatively, natural phenotypes not involving artificial processing were more directly controlled by genetic factors (Melchinger et al. 2004). Traits with high correlation coefficients clustered within the same linkage group (e.g., agronomic traits clustered in LG6), while those with low correlation coefficients were distributed across different LGs. These results indicate that the genetic map/SNP markers are effective for additional QTL studies and provide reference information for further QTL mapping of complex traits.

We also performed a pilot test of GS in tobacco. This approach is a more efficient strategy to predict and select for complex traits than marker-assisted selection, which is usually based on established mapping results (Crossa et al. 2017; Yamamoto et al. 2017; Cerrudo et al. 2018). The prediction accuracy was mainly determined by the population size and proportion (ratio of the test set to the training set) and was weakly related to the density of markers. The genomic prediction analysis was conducted on the RIL population using different marker numbers, population sizes, and models. Population size was most effective factor for improving the predication accuracy and the BayesB model may be the most well-suited for the tobacco RIL population. Combined, our results demonstrate that GS prediction is promising in this tobacco RIL population, supporting the application of this approach to tobacco breeding.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11816-021-00713-1.

Acknowledgements This work was supported by the Fundamental Research Program of Yunnan Province (2018FB064), China National Tobacco Company (110201901014 (JY-01), no. 110201801029(JY-06)) and Yunnan Tobacco Company (nos. 2020530000241009, 2020530000241007, 2018530000241005).

Author contributions BX and JJ conceived the research and designed the experiments, ZT, DF, FJ, and XC developed the RIL population and collected phenotypic data, YM, YH, CZ and JZ performed statistical analysis of the data, ZT, ZX and JJ wrote the manuscript. HJ, YM, BX WH and SZ revised the manuscript. All authors reviewed and approved this submission.

#### Declarations

Conflict of interest The authors declare no conflict of interest.

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

#### References

- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
- Bao F, Du D, An Y, Yang W, Wang J, Cheng T, Zhang Q (2017) Overexpression of Prunus mume dehydrin genes in tobacco enhances tolerance to cold and drought. Front Plant Sci 8:151
- Bates D, Sarkar D, Bates MD et al (2007) The lme4 package [J]. R package version 2(1):74
- Bennewitz S, Bergau N, Tissier A (2018) QTL mapping of the shape of type VI glandular trichomes in tomato. Front Plant Sci 9:1421
- Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V, Tyagi A, Mushtaq M, Jain N, Singh PK, Singh GP, Prabhu KV (2016) Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Front Genet 7:221
- Cerrudo D, Cao S, Yuan Y, Martinez C, Suarez EA, Babu R, Zhang X, Trachsel S (2018) Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. Front Plant Sci 9:366
- Cheng L, Yang A, Jiang C, Ren M, Zhang Y, Feng Q, Wang S, Guan Y, Luo C (2015) Quantitative trait loci mapping for plant height in tobacco using linkage and association mapping methods. Crop Sci 55:641–647
- Cheng L, Chen X, Jiang C, Ma B, Ren M, Cheng Y, Liu D, Geng R, Yang A (2019) High-density SNP genetic linkage map construction and quantitative trait locus mapping for resistance to cucumber mosaic virus in tobacco (*Nicotiana tabacum* L.). Crop J 7:539–547
- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971
- Crossa J, Perez-Rodriguez P, Cuevas J, Montesinos-Lopez O, Jarquin D, de Los Campos G, Burgueno J, Gonzalez-Camacho JM, Perez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975
- Diouf IA, Derivot L, Bitton F, Pascual L, Causse M (2018) Water deficit and salinity stress reveal many specific QTL for plant growth and fruit quality traits in tomato. Front Plant Sci 9:279
- Dixit S, Singh A, Sandhu N, Bhandari A, Vikram P, Kumar A (2017) Combining drought and submergence tolerance in rice: markerassisted breeding and QTL combination effects. Mol Breed 37:143
- Drake-Stowe K, Bakaher N, Goepfert S, Philippon B, Mark R, Peterson P, Lewis RS (2017) Multiple disease resistance loci affect soilborne disease resistane in Tobacco (*Nicotiana tabacum*). Phytopathology 107:1055–1061
- Edwards KD, Bombarely A, Story GW, Allen F, Mueller LA, Coates SA, Jones L (2010) TobEA: an atlas of tobacco gene expression from seed to senescence. BMC Genom 11:142
- Endelman JB (2011) Ridge regression and other kernels for genomic selection with R Package rrBLUP. Plant Genome 4:250–255
- Ganal MW, Altmann T, Roder MS (2009) SNP identification in crop plants. Curr Opin Plant Biol 12:211–217
- Hoekema A, Hirsch PR, Hooykaas PJJ, Schilperoot RA (1983) A binary plant vector strategy based on separation of vir- and T-region of the agrobacterium tumefaciens Ti-plasmid. Nature 303:179–180
- Hu Z, Lu SJ, Wang MJ, He H, Sun L, Wang H, Liu XH, Jiang L, Sun JL, Xin X, Kong W, Chu C, Xue HW, Yang J, Luo X, Liu JX (2018) A novel QTL qTGW3 encodes the GSK3/SHAGGY-like

kinase OsGSK5/OsSK41 that interacts with OsARF4 to negatively regulate grain size and weight in rice. Mol Plant 11:736–749

- Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics 135:205–211
- Julio E, Denoyes-Rothan B, Verrier JL, de Borne F (2006) Detection of QTLs linked to leaf and smoke properties in *Nicotiana tabacum* based on a study of 114 recombinant inbred lines. Mol Breed 18:69–91
- Kenton A, Parokonny AS, Gleba YY, Bennett MD (1993) Characterization of the *Nicotiana tabacum* L. genome by molecular cytogenetics. Mol Gen Genet 240:159–169
- Lee YK, Kim IJ (2018) Functional conservation of Arabidopsis LNG1 in tobacco relating to leaf shape change by increasing longitudinal cell elongation by overexpression. Genes Genom 40:1053–1062
- Leitch IJ, Hanson L, Lim KY, Kovarik A, Chase MW, Clarkson JJ, Leitch AR (2008) The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). Ann Bot 101:805–814
- Lewis RS, Milla SR, Kernodle SP (2007) Analysis of an introgressed Nicotiana tomentosa genomic region affecting leaf number and correlated traits in *Nicotiana tabacum*. TAG Theor Appl Genet 114:841–854
- Liu YJ, Xiu ZH, Meeley R, Tan BC (2013) Empty pericarp5 encodes a pentatricopeptide repeat protein that is required for mitochondrial RNA editing and seed development in maize. Plant Cell 25:868–883
- Liu Y, Wang L, Sun C, Zhang Z, Zheng Y, Qiu F (2014) Genetic analysis and major QTL detection for maize kernel size and weight in multi-environments. TAG Theor Appl Genet 127:1019–1037
- Ma JM, Heim C, Humphry M, Nifong JM, Lewis RS (2019) Genetic analysis of Phn7.1, a major QTL conferring partial resistance to *Phytophthora nicotianae* in *Nicotiana tabacum*. Mol Breed 39:11
- McAdam K, Eldridge A, Fearon IM, Liu C, Manson A, Murphy J, Porter A (2016) Influence of cigarette circumference on smoke chemistry, biological activity, and smoking behaviour. Regul Toxicol Pharmacol 82:111–126
- Melchinger AE, Utz HF, Schon CC (2004) QTL analyses of complex traits with cross validation, bootstrapping and other biometric methods. Euphytic 137:1–11
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829
- Misztal I, Legarra A (2017) Invited review: efficient computation strategies in genomic selection. Anim Int J Anim Biosci 11:731–736
- Mulder N, Apweiler R (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 396:59–70
- Nocarova E, Fischer L (2009) Cloning of transgenic tobacco BY-2 cells; an efficient method to analyse and reduce high natural heterogeneity of transgene expression. BMC Plant Biol 9:44
- Pagliano C, Bersanini L, Cella R, Longoni P, Pantaleoni L, Dass A, Leelavathi S, Reddy VS (2017) Use of *Nicotiana tabacum* transplastomic plants engineered to express a His-tagged CP47 for the isolation of functional photosystem II core complexes. Plant Physiol Biochem PPB 111:266–273
- Prat N, Guilbert C, Prah U, Wachter E, Steiner B, Langin T, Robert O, Buerstmayr H (2017) QTL mapping of Fusarium head blight resistance in three related durum wheat populations. TAG Theor Appl Genet 130:13–27
- Robertsen C, Hjortshøj R, Janss L (2019) Genomic selection in cereal breeding. Agronomy 9:95
- Rushton PJ, Bokowiec MT, Laudeman TW, Brannock JF, Chen X, Timko MP (2008) TOBFAC: the database of tobacco transcription factors. BMC Bioinform 9:53
- Sierro N, Battey JN, Ouadi S, Bovet L, Goepfert S, Bakaher N, Peitsch MC, Ivanov NV (2013) Reference genomes and transcriptomes

of Nicotiana sylvestris and Nicotiana tomentosiformis. Genome Biol 14:R60

- Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV (2014) The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun 5:3833
- Sun F, Wang X, Bonnard G, Shen Y, Xiu Z, Li X, Gao D, Zhang Z, Tan BC (2015) Empty pericarp7 encodes a mitochondrial E-subgroup pentatricopeptide repeat protein that is required for ccmFN editing, mitochondrial function and seed development in maize. Plant J 84:283–295
- Thimmegowda GC, Ramadoss SK, Kaikala V, Rathinavelu R, Thamalampudi VR, Dhavala VNC, Saiprasad GVS (2018) Whole genome resequencing of tobacco (*Nicotiana tabacum* L.) genotypes and high-throughput SNP discovery. Mol Breed 38:121
- Tong Z, Jiao T, Wang F, Li M, Leng X, Gao Y, Li Y, Xiao B, Wu W (2012) Mapping of quantitative trait loci conferring resistance to brown spot in flue-cured tobacco (*Nicotiana tabacum* L.). Plant Breed 131:335–339
- Tong Z, Zhou J, Xiu Z, Jiao F, Hu Y, Zheng F, Chen X, Li Y, Fang D, Li S, Wu X, Zeng J, Zhao S, Jian J, Xiao B (2020) Construction of a high-density genetic map with whole genome sequencing in *Nicotiana tabacum* L. Genomics 112:2028–2033
- Tura H, Edwards J, Gahlaut V, Garcia M, Sznajder B, Baumann U, Shahinnia F, Reynolds M, Langridge P, Balyan HS, Gupta PK, Schnurbusch T, Fleury D (2020) QTL analysis and fine mapping of a QTL for yield-related traits in wheat grown in dry and hot environments. TAG Theor Appl Genet 133:239–257
- Vasil V, Hildebrandt AC (1965) Differentiation of tobacco plants from single, isolated cells in microcultures. Science 150:889–892
- Vinocur B, Altman A (2005) Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations. Curr Opin Biotechnol 16:123–132
- Wang GF, Balint-Kurti PJ (2016) Maize homologs of CCoAOMT and HCT, two key enzymes in lignin biosynthesis, form complexes with the NLR Rp1 Protein to modulate the defense response. Plant Physiol 171:2166–2177
- Wang X, Bennetzen JL (2015) Current status and prospects for the study of Nicotiana genomics, genetics, and nicotine biosynthesis genes. Mol Genet Genom MGG 290:11–21
- Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: current status and prospects. Crop J 6:330–340
- Wu Q, Wu X, Zhang X, Jiang C, Xiao B, Zhang Y, Wang Y, Liu G (2014) Mapping of two white stem genes in tetraploid common tobacco (*Nicotiana tabacum* L.). Mol Breed 34:1065–1074
- Xiao B, Tan Y, Long N, Chen X, Tong Z, Dong Y, Li Y (2015) SNPbased genetic linkage map of tobacco (*Nicotiana tabacum* L.) using next-generation RAD sequencing. J Biol Res (thessalon) 22:11
- Xiu Z, Sun F, Shen Y, Zhang X, Jiang R, Bonnard G, Zhang J, Tan BC (2016) EMPTY PERICARP16 is required for mitochondrial nad2 intron 4 cis-splicing, complex I assembly and seed development in maize. Plant J 85:507–519
- Yamamoto E, Matsunaga H, Onogi A, Ohyama A, Miyatake K, Yamaguchi H, Nunome T, Iwata H, Fukuoka H (2017) Efficiency of genomic selection for breeding population design and phenotype prediction in tomato. Heredity 118:202–209
- Zhang X, Huang C, Wu D, Qiao F, Li W, Duan L, Wang K, Xiao Y, Chen G, Liu Q, Xiong L, Yang W, Yan J (2017) High-throughput phenotyping and QTL mapping reveals the genetic architecture of maize plant growth. Plant Physiol 173:1554–1564
- Zhang Y, Guo X, Yan X, Ren M, Jiang C, Cheng Y, Wen L, Liu D, Zhang Y, Sun M, Feng Q, Yang A, Cheng L (2018) Identification of stably expressed QTL for resistance to black shank disease in tobacco (*Nicotiana tabacum* L.) line Beinhart 1000–1. Crop J 6:282–290

Zhang H, Yin L, Wang M, Yuan X, Liu X (2019) Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front Genet 10:189

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