Identification and analysis of novel recessive alleles for *Tan1* and *Tan2* in sorghum (#84664)

First submission

Guidance from your Editor

Please submit by **11 Jun 2023** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files

Download and review all files from the materials page.

6 Figure file(s)
10 Table file(s)

Custom checks

DNA data checks

- Have you checked the authors data deposition statement?
- Can you access the deposited data?
- Has the data been deposited correctly?
- Is the deposition information noted in the manuscript?

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Identification and analysis of novel recessive alleles for *Tan1* and *Tan2* in sorghum

Lixia Zhang [1], **Chunyu Wang** [1], **Miao Yu** [2], **Ling Cong** [1], **Zhenxing Zhu** [1], **Bingru Chen** [Corresp., 2], **Xiaochun Lu** [Corresp. 1]

[1] Sorghum Research Institute, Liaoning Academy of Agricultural Sciences, Shenyang, Shenhe, China

[2] Institute of Crop Germplasm Resources, Jilin Academy of Agricultural Sciences, Gongzhuling, Kemaoxi Street, China

Corresponding Authors: Bingru Chen, Xiaochun Lu
Email address: chenbingru1979@163.com, luxiaochun2000@126.com

Background: The identification and analysis of allelic variation are important bases for crop diversity research, trait domestication and molecular marker development. Grain tannin content is a very important quality-related trait in sorghum. Higher tannin levels in sorghum grains are usually required when breeding varieties resistant to bird damage or those used for brewing liquor. Non-tannin-producing or low-tannin-producing sorghums are commonly used for food and forage. *Tan1* and *Tan2*, two important cloned genes, regulate tannin biosynthesis in sorghum, and mutations in one or two genes will result in low or no tannin content in sorghum grains. Even if sorghum accessions contain *Tan1Tan2*, the tannin contents are distributed from low to high, and there must be other new alleles or new regulatory genes. **Methods:** The two parents (8R306/8R191) had *Tan1Tan2* genotype and tannins and nontannins in the grains, were constructed a RIL population. BSA (Bulked Segregant Analysis) was used to determine the new Tannin locus. *Tan1* and *Tan2* full-length sequences and tannin contents were detected in landraces and cultivars. **Results:** We identified two novel recessive *tan1-d* and *tan1-e* alleles and four recessive *tan2* alleles named *tan2-d*, *tan2-e*, *tan2-f* and *tan2-g*. All these recessive alleles lead to loss of function of Tan1 and Tan2, and low or no tannin content in sorghum grains. The loss-of-function alleles of *tan1-e* and *tan2-e* were only found in Chinese landraces, and other alleles were found in landraces and cultivars grown all around the world. *tan1-a* and *tan1-b* were detected in domestic and foreign sorghum cultivars and foreign landraces but not in Chinese landraces. **Conclusion:** These results imply that *Tan1* and *Tan2* genes have undergone different evolutionary trajectories in different planting areas worldwide, and not all *tan1*and *tan2* alleles have been used in breeding. Discovery of these new alleles provided new germplasm resources for breeding sorghum cultivars for food and feed and for developing molecular markers for low-tannin cultivar-assisted breeding in sorghum.

Identification and analysis of novel recessive alleles for *Tan1* and *Tan2* in sorghum

Lixia Zhang[1], Chunyu Wang[1], Miao Yu[2], Ling Cong[1], Zhenxing Zhu[1], Bingru Chen[2, *], Xiaochun Lu[1, *]

[1] Sorghum Research Institute, Liaoning Academy of Agricultural Sciences, Shenyang, Liaoning Province, China

[2] Institute of Crop Germplasm Resources, Jilin Academy of Agricultural Sciences, Gongzhuling, Jilin Province, China

Corresponding Author:
Xiaochun Lu[1], Dongling Stree, Shenyang, Liaoning Province, 110161, China
Email address: luxiaochun2000@126.com
Bingru Chen[2], Kemaoxi Street, Gongzhuling, Jilin Province, 136100, China
Email address: chenbingru1979@163.com

Abstract

Background: The identification and analysis of allelic variation are important bases for crop diversity research, trait domestication and molecular marker development. Grain tannin content is a very important quality-related trait in sorghum. Higher tannin levels in sorghum grains are usually required when breeding varieties resistant to bird damage or those used for brewing liquor. Non-tannin-producing or low-tannin-producing sorghums are commonly used for food and forage. *Tan1* and *Tan2*, two important cloned genes, regulate tannin biosynthesis in sorghum, and mutations in one or two genes will result in low or no tannin content in sorghum grains. Even if sorghum accessions contain *Tan1Tan2*, the tannin contents are distributed from low to high, and there must be other new alleles or new regulatory genes.

Methods: The two parents 8R306 and 8R191) had *Tan1Tan2* genotype and tannins and nontannins in the grains, were constructed RIL population. BSA (Bulked Segregant Analysis) was used to determine the new Tannin locus. *Tan1* and *Tan2* full-length sequences and tannin contents were detected in landraces and cultivars.

Results: We identified two novel recessive *tan1-d* and *tan1-e* alleles and four recessive *tan2* alleles named *tan2-d*, *tan2-e*, *tan2-f* and *tan2-g*. All these recessive alleles lead to loss of

function of Tan1 and Tan2, and low or no tannin content in sorghum grains. The loss-of-function alleles of *tan1-e* and *tan2-e* were only found in Chinese landraces, and other alleles were found in landraces and cultivars grown all around the world. *tan1-a* and *tan1-b* were detected in domestic and foreign sorghum cultivars and foreign landraces but not in Chinese landraces. **Conclusion:** These results imply that *Tan1* and *Tan2* genes have undergone different evolutionary trajectories in different planting areas worldwide and not all *tan1* and *tan2* alleles have been used in breeding. Discovery of these new alleles provided new germplasm resources for breeding sorghum cultivars for food and feed and for developing molecular markers for low-tannin or non-tannin cultivar-assisted breeding in sorghum.

Keywords Sorghum bicolor, Tannins, Allele, Domestication

Introduction

Sorghum [*S. bicolor* (L.) Moench] is the fifth largest food crop in the world and is widely used for producing food, feed, brewed beverages and biofuel (Dahlberg 2019; Zhao et al. 2019). Tannins (also known as condensed tannins or proanthocyanidins) are important for the perception of quality in sorghum, and the tannin content determines the use of sorghum grains. Tannins are widespread in fruits, nuts, vegetables and some cereals (He et al. 2008). During crop domestication and evolutionary processes, tannin production was removed from major cereal crops (such as rice, wheat, and maize) but was retained in finger millet, barley and sorghum (Zhu 2019). Tannins, with diverse biological and biochemical functions, have negative impacts on nutritional value, such as decreasing protein digestibility and feed efficiency in humans and animals (Choi & Kim 2020; Chung et al. 1998). Therefore, non-tannin-producing or low-tannin-producing sorghum cultivars are used in food and feeding production. However, tannins can promote human health because of their high antioxidant capacity and ability to fight obesity through reduced digestion (Cos et al. 2004; Habyarimana et al. 2019). Tannin-producing and non-tannin-producing (or low-tannin-producing) sorghum cultivars are widely grown worldwide for their different applications and economic values. Evaluating the data on the presence of tannins in 11,557 cultivated sorghum accessions in Africa, approximately 55% were of the non-tannin-producing type and 45% were of the tannin-producing type (Wu et al. 2019). In China, high-tannin-producing sorghum is particularly important in liquor production, accounting for 80% of China's total sorghum production. The well-known Moutaijiu, Langjiu, Luzhoulaojiao, Wuliangye and several other famous liquors are fermented by using high-tannin-producing

sorghum as a main feedstock (Zhang et al. 2022). The coexistence of tannin-producing and non-tannin-producing (or low-tannin-producing) sorghum suggests that the elimination of this compound from sorghum grains during domestication was incomplete, exemplifying strong artificial selection against tannins in breeding and production.

Tannins (proanthocyanidins) and anthocyanins are major flavonoid end-products from a well-conserved family of aromatic molecules that have several biological functions in plant development and defense (Gutierrez et al. 2020; Huang et al. 2019; Xie et al. 2019; Xie & Xu 2019). Tannins are derived from a branch of the flavonoid pathway, as well documented in Arabidopsis. AtTT2/AtTT8/AtTTG1 forms an MBW complex (MYB-bHLH-WD40) to regulate tannin synthesis (Baudry et al. 2004; Li et al. 2020; Schaart et al. 2013; Wang et al. 2017). Using genetic linkage mapping, *Tannin1* (*Tan1*, Sobic.004G280800) and *Tannin2* (*Tan2*, Sobic.002G076600) have been cloned in sorghum (Wu et al. 2019; Wu et al. 2012). Tan1 encodes a WD-40 repeat protein, and Tan2 encodes a bHLH domain protein, both have a regulatory function similar to that of Arabidopsis AtTTG1 and AtTT8. Three loss-of-function alleles each for *Tan1* and *Tan2* were identified in sorghum, including *tan1-a*, *tan1-b*, and *tan1-c* and *tan2-a*, *tan2-b* and *tan2-c*. Low or no tannins in sorghum grains can result from recessive alleles at one or both of these loci. A genome-wide association study (GWAS) was used to detect other tannin-related loci to identify more loci underlying natural variation in grain tannin content and pigmentation. Three highly significant association peaks spanning were observed, including 1.16-1.23 Mb (Chr1), 8.075-8.45 Mb (Chr2) and 57.9 Mb (Chr3), suggesting that other genes controlling this trait may exist (Morris et al. 2013). Discovery of new tannin-regulating genes will provide a better and more accurate detection method for cultivating special sorghum varieties with different tannin contents.

To develop practical molecular markers for tannin breeding, more *tan1* and *tan2* alleles need to be detected. We used wild sorghum, as well as landraces and cultivars, to comprehensively identify the alleles of *tan1* and *tan2*. We identified two novel recessive *tan1* alleles and four recessive *tan2* alleles by map-based cloning and sequencing *Tan1* and *Tan2* coding regions. These new alleles will provide a solid foundation to study the evolution of *Tan1* and *Tan2* and their artificial selection in cultivar breeding and provide genetic resources for breeding non-tannin-producing or low-tannin-producing sorghum cultivars.

Materials & Methods

Plant materials

Sorghum accessions include wild sorghums, landraces and cultivars from all over the world that were collected from the Sorghum Institute, Liaoning Academy of Agricultural Sciences, China. Plants were grown at the experimental site of Liaoning Academy of Agricultural Sciences. Leaf tissue was collected, frozen in liquid nitrogen and stored at -80 °C until further use. Grains were harvested to determine the tannin contents.

DNA extraction

Leaves from each sorghum accession were sampled for genomic DNA extraction by the cetyltrimethylammonium bromide (CTAB) method as previously described with minor modifications (Allen et al. 2006).

PCR, DNA sequencing, and sequence analysis

To genotype *Tan1* and *Tan2* alleles in different sorghum accessions, primers were designed (Table S1). The PCR products were sequenced by Beijing Tsingke Biological Technology Co., Ltd. (Beijing, China). The DNAMAN program (version 5.2.2) was used for sequence alignment and translation of nucleotides into amino acids. To develop the CAPS (cleaved amplified polymorphic sequence) marker to detect the *tan1-c* allele, Tan1-2F/ Tan1-2R combined with *Dde* I was designed for *tan1-c* (Table S1). PCR products were digested with *Dde* I and analyzed by 8% polyacrylamide gel electrophoresis. Because *tan1-c* lost a *Dde* I restriction enzyme site as a result of A-to-T transversion at position 1054 in the coding sequence, PCR amplification with Tan1-2F/ Tan1-2R resulted in a 162 bp product; whereas *Tan1* contained a single *Dde* I site in the corresponding PCR product and was cut into 109 bp and 56 bp fragments.

Determination of tannin content by reagent test kit

Tannin was determined according to the Tannin Microplate Assay Kit (Cohesion Biosciences, CAK1060). Five grams of grains were crushed into a powder in a grinder. Tissue samples (0.1 g) were homogenized with 1 ml of distilled water, placed in a water bath at 80 °C for 30 minutes, and centrifuged at 8,000 g at 4 °C for 10 minutes. The supernatant was placed into a new centrifuge tube for detection. Ten microliters of sample supernatant, 160 µl of distilled water and 20 µl of reaction buffer were mixed and incubated for 5 minutes at room temperature. Then, 10 µl of dye reagent was mixed for 10 minutes, and the absorbance was measured and recorded at 650 nm to calculate the tannin content. The tannin contents were

scored as low (≤0.5%), medium (0.5%>tannin content<1.0%) and high (≥1%). In this study, medium tannin content accessions were not exhibited.

Mapping Population

8R191 was a Chinese landrace without tannins. 8R306 was a landrace from African and had tannins in the grain. The RIL populations were obtained by single-seed descent from 8R306 and 8R109 with 557 lines.

Mapping and identification of the candidate gene

Genomic DNAs were extracted from 8R191, 8R306 and RIL population plants using the CTAB method. BSA (Bulked Segregant Analysis) was used to determine the new Tannin locus. Equal amounts of genomic DNAs from 50 tannin and 50 non-tannin plants were pooled to construct the tannin and non-tannin bulks, respectively.

High-throughput genome sequencing and data analysis of the two bulks and two parental lines were conducted by Beijing PlantTech Biotechnology Co., Ltd (Beijing, China). ΔSNP method was applied to associate the new Tannin locus using Sorghum bicolor v3.1.1 as the reference genome (https://phytozome-next.jgi.doe.gov/info/Sbicolor_v3_1_1). InDel and SNP markers within the region were used for fine mapping. Information on molecular markers for fine mapping is provided in Supplementary Data Table S1.

Chlorox Bleach Test

Chlorox bleach test was performed previously described with minor modifications (Dykes 2019). Put 100 sorghum grains into a 100 ml beaker, add15 ml 6% NaClO to fully immerse sorghum grains. Sit the beaker for 20 min at room temperature and swirl the contents in the beaker every 5 min. Discard the reaction solution, rinse it with distilled water 2-3 times, and pour it on filter paper to remove excess water. All bleach tests were repeated three times. The presence or absence of tannins in sorghum grains was evaluated based on grain color after dyeing. Sorghum grains are divided into three types, Type I grains were completely black and had tannins, Type II grains were lighter brown black or had small black spots, and Type III grains were white or lightly colored and had no tannins. In 557 RILs, Type I had 109 lines, Type II had 179 lines and Type III had 269 lines. For the accuracy of phenotypic identification, Type I and Type III lines were used to map the new *Tannin* locus. After dyeing, select one grain from one line to use for germinating and sampling the seedling to extract the genomic DNAs.

Results

Relationships between *Tan1* and *Tan2* genotypes and tannin contents

The presence of tannins in sorghum grains is regulated by a pair of genes (*Tan1* and *Tan2*), and both genes have three recessive alleles (Wu et al. 2019; Wu et al. 2012). Twenty accessions were used to determine the relationship between *Tan1* and *Tan2* genotypes and tannin contents in sorghum grains. As shown in Table 1, homozygous recessive genotypes at one or both genes can cause a low-tannin (or non-tannin) phenotype, and two wild sorghum accessions had high tannin contents because they carry dominant alleles, which was consistent with reported data (Wu et al. 2019). However, some accessions carrying *Tan1* and *Tan2* dominant alleles had low tannin contents in sorghum grains, indicating that they may have new genes or new alleles for two known genes (Table 1).

A novel tannin recessive allele of *Tan1* in sorghum landraces

To identify new tannin genes or new alleles, a recombinant inbred line (RIL) population was built from 8R191/8R306. Sequencing and digestion were used to determinate the *Tan1* and *Tan2* alleles. Amplifying and sequencing by Tan1-1F/ Tan1-1R indicated that 8R191 and 8R306 didn't have *tan1-a* and *tan1-b* (Table S1). A CAPS marker was designed to detect the *tan1-c* allele. 165 bp PCR products with Tan1-2F/2R primers were digested by *Dde* I, 109 bp and 56 bp DNA fragments of *Tan1*. The PCR product of 8R191 and 8R306 can cleave indicating that two parents didn't have *tan1-c*. Because of A-to-T transversion at position 1054 in the coding sequence of *tan1-c*, the PCR products of Tx2752, OK11 and Tx631 remained uncleaved (Table S1, Figure 1) (Wu et al. 2019). 6 primers were used to detect *Tan2* genotypes in 8R191 and 8R306 (Table S1). 8R191 is a non-tannin landrace and 8R306 is a tannin landrace, both of them carrying *Tan1* and *Tan2* dominant alleles.

We performed BSA using the 8R191/8R306 RIL population. Equal amounts of genomic DNAs from 50 tannin and 50 non-tannin plants were pooled to construct the tannin and non-tannin bulks, respectively. The 8R191(non-tannin parent), 8R306(tannin parent), and non-tannin, tannin bulks were subjected to Illumina high-throughput sequencing, from which, 70.99, 72.90, 290.71 and 280.82 million paired-end reads were produced, representing 14×, 14×, 54×, and 53× genome coverage, respectively (Table S2). Among them, 98.23%, 97.88%, 94.35% and 95.91% reads could be mapped to the reference genome, respectively indicating good quality of the sequencing data (Table S2). Using BSA-Seq method, we obtained only one region spanning 6.60

Mb on Chr4 between 57,400,000 and 64,000,000 that was strongly associated with the tannin phenotype (Figure 2A). Within this region, we developed 10 available InDel markers for fine mapping (Figure 2B). Using a RIL population of 378 plants (Type I: 109 lines and Type III: 269 lines), the new *Tannin* locus was finally narrowed down to a 25.8kb region defined by markers InDel-7 and InDel-8. We identified 4 candidate genes in this region, including Sobic.004G280700, Sobic.004G280800, Sobic.004G280900 and Sobic.004G281000 (Figure 2C, Table S3). Among these genes, Sobic.004G280800 is *Tan1*, suggesting that Sobic.004G280800 is likely to be the causal gene. Primers (Tan1-F/Tan1-R, Table S1) were used to detect the sequence polymorphisms in *Tan1* between 8R191 and 8R306. 8R306 had the *Tan1* allele, however, 8R191 had deletion and subsititution in coding region of *Tan1*, and named *tan1-d*. Therefore, the *Tan1* has abundant allelic variation genotypes.

Identification and distribution of *Tan1* and *Tan2* other new alleles

To identify more different *Tan1* and *Tan2* alleles, we collected diverse sorghum accessions, including wild sorghum, landraces and cultivars (Table S4). Another novel *tan1* recessive allele was found, named *tan1-e*. Among 396 sorghum accessions, 10 wild sorghum and 200 high-tannin-producing accessions carried the *Tan1* allele, and 89 low-tannin-producing accessions carried the *tan1-a*, *tan1-b*, *tan1-d and tan1-e* alleles. A total of 97 accessions carried the *Tan1* allele but had low tannin contents (Table 2). Seventy-two accessions carrying the *Tan1* allele were used to detect different *Tan2* alleles, including 38 low-tannin-producing accessions, 24 high-tannin-producing accessions and 10 wild sorghum accessions. The ten wild sorghum accessions and 24 high-tannin-producing accessions had dominant *Tan1* and *Tan2* alleles (Table S5). Four different recessive alleles of *Tan2* were identified, named *tan2-d*, *tan2-e*, *tan2-f* and *tan2-g* (Table 3, Table S5). More importantly, the *tan1* and *tan2* recessive alleles had obvious regional distribution characteristics. Novel *tan1-d*, *tan2-d*, *tan2-f* and *tan2-g* alleles were distributed worldwide, including Ghana, France, Mexico, India, China and so on, but *tan1-e* and *tan2-e* were only found in Chinese landraces (Table S4 and S5). The results implied that *Tan1* and *Tan2* may have been selected artificially after independent evolution in ecotype areas.

Functional variation of the newly identified *Tan1* and *Tan2* alleles

Tan1 and *Tan2* are conserved regulatory factors in the plant tannin synthesis pathway and have higher nucleotide similarity within major cereal crops other than rice, wheat and maize,

which do not produce tannins in their grains. Tan1 encodes a WD-40 repeat protein and has four WD-40 repeat domains. Deletion, substitution and insertion mutations in *tan1-a*, *tan1-b* and *tan1-c* have caused frame shifts and premature stop codons, leading to disruption of the highly conserved region of the WD-40 domain and C-terminus and resulting in the absence or low level of tannins in sorghum grains. Seven independent mutations of the *TTG1* gene revealed that the truncation of the C-terminal region and WD-40 domain produced nonfunctional alleles in Arabidopsis, indicating that the C-terminal region and WD-40 domain are vital for the structure and function of the WD-40 protein (Wu et al. 2012). In *tan1-d*, A-to-T transversion at position 1054, GT deletion at position 1057 and 1058, and C-to-T transition at position 1059 in the coding sequence affected TGA (at position 1060, 1061 and 1062) stop codon frameshift and led to nonfunctional protein. Because of mutation and deletion, *tan1-d* had 1086 bp coding sequence and 361 aa protein sequence (Figure 3A and 3B). Sequence variation of *tan1-d* was similar to *tan1-c*, the C-terminal sequence had changed greatly (Figure 4). Compared to dominant *Tan1*, *tan1-e* has a 10-bp deletion (TCGACATACG) in the coding sequence between positions 771 and 780. The 10-bp deletion causes a frameshift and results in a truncated protein with a length of only 295 aa (Figure 3A and 3B). The 10-bp deletion in *tan1-e* also results in shifts in the third and fourth WD-40 repeat domains and C-terminal region, similar to *tan1-a* and *tan1-b*.

Tan2 encodes a bHLH transcription factor with 10 exons and 9 introns in BTx623. *tan2-a* has a 5-bp (CCCCT) insertion in the 8th exon, *tan2-b* has a 7-bp (AGACCAC) insertion in the 7th exon and *tan2-c* has a 95-bp deletion removing the entire 8th intron. These mutations lead to file:///C:/Users/Administrator.USER-20180330PS/Desktop/Peer J-Tannin/figures and tables/figures and tables/Figure6.pngframe shifts, disrupt the bHLH domain and result in the non-tannin-producing or low-tannin-producing phenotype (Wu et al. 2019). In our study, the A-to-G transition at position 51, T-to-G transversion at position 1302, T-to-C transition at position 1428 and G-to-C transversion at position 1569 in the coding sequence did not affect protein function because of synonymous mutations and agreed with the reported data (Figure 5, Figure 6) (Wu et al. 2019). *tan2-d*, with the causal polymorphism of a 1-bp C deletion at position 563 in the coding region, leads to a truncated protein with a length of only 210 aa. Because of the C-to-T transition at position 1366 (**C**AG to **T**AG) in the coding sequence, *tan2-e* results in premature termination and a 455 aa protein. *tan2-f* contains a frameshift and an early termination site because of an 8-bp (AGCTGATC) insertion between positions 1375 and 1376 in the coding

region, resulting in a 462 aa protein sequence. *tan2-g* has multiple subsitons and insertions from positions 1579 to 1607 in the coding region that have led to disruption of the bHLH domain structure (Figure 5, Figure 6).

***Tan1* and *Tan2* allele utilization in breeding programs**

By investigating the distribution of *Tan1* and *Tan2* alleles in sorghum cultivars, we can determine which alleles have been used in breeding. 87 cultivars (from China and foreign countries), as well as 34 sterile lines and 43 restorer lines, were used to detect the different alleles of *Tan1* and *Tan2* (Table S4 and S6). For *Tan1*, only the *tan1-a*, *tan1-b* and *tan1-c* alleles were detected in Chinese sorghum cultivars, neither *tan1-d* nor *tan1-e* were found, which means that low-tannin-producing resources that are used in breeding have been mainly introduced from India, the USA and other countries and that *tan1-e* (only detected in Chinese low-tannin-producing landraces) may not have been adopted (Table S4 and S6). For *Tan2*, *tan2-a*, *tan2-b* and *tan2-c* were detected in cultivars. In our study, *tan2-f* and *tan2-g* alleles were detected in cultivars, and *tan2-d* and *tan2-e* alleles were not (Table S5). More importantly, *tan2-e* was only detected in Chinese landraces (Table S5). The results showed that only a few kinds of *tan1* and *tan2* alleles were applied in breeding, which will lead to reduced diversity in breeding resources.

Discussion

Wild sorghum accessions generally show higher tannin contents than domesticated accessions due to selection during domestication (Dykes & Rooney 2007). The apparent nutrient absorption and protein digestion issues were reduced by feeding sorghum grains with high tannin content. Sorghum breeding programs mainly rely on grain color to determine the contents of tannins in grains (Rhodes et al. 2014). Sorghums with pigmented testa contain concentrated tannins. The use of grain color as a proxy for tannin concentration is complicated by the need for varietal information, including pigmented testa and endosperm appearance, which are correlated with tannin levels (Dykes 2019; Oliveira et al. 2017). Using marker-assisted breeding can simplify and expedite breeding for determining the tannin content. Identifying tannin-related genes and alleles is very important for molecular selection and breeding.

AtTT2, AtTT8 and AtTTG1 form an MBW complex to regulate tannin synthesis (Baudry et al. 2004; Ha et al. 2018; Schaart et al. 2013). Nonfunctional AtTTG1 and AtTT8 proteins impact MBW complex function, which inhibits the expression of *DFR*, *LAR* and *ANR* and hinders tannin

synthesis (Shan et al. 2019; Sun et al. 2022; Wei et al. 2019). *Tan1* (homologous gene-*AtTTG1*) and *Tan2* (homologous gene-*AtTT8*) are involved in regulating the tannin synthesis pathway in sorghum, and three recessive alleles each for *tan1* and *tan2* have been reported (Wu et al. 2019; Wu et al. 2012). In our study, two novel recessive alleles for *Tan1* and four novel recessive alleles for *Tan2* were identified, including *tan1-d*, *tan1-e*, *tan2-d*, *tan2-e*, *tan2-f* and *tan2-g* (Figs 3 and 5, Table S7). Because the insertion or deletion positions in the coding regions of the five recessive *tan1* alleles and seven recessive *tan2* alleles are different, their corresponding Tan1 and Tan2 proteins are nonfunctional but show variable inhibition of tannin accumulation in sorghum grains. These alleles will be useful for marker-assisted breeding for the improvement of low-tannin-producing or non-tannin-producing sorghum cultivars.

The tannin contents of 186 out of 396 accessions were under 0.5%. These accessions were widely distributed in China, India, Africa and other countries, consistent with their high number of recessive alleles for tannin synthesis-regulating genes (Table S4 and S5). These low-tannin-producing accessions may contain the five recessive *tan1,* seven recessive *tan2* alleles or the dominant *Tan1* and *Tan2* alleles, indicating that there are unknown alleles or tannin-regulated genes. The identified *tan1* and *tan2* alleles have certain characteristics of regional distribution; for example, *tan1-e* and *tan2-e* are only distributed in China (Table S4 and S5).

There are many different characteristics among Chinese sorghums, African sorghums and Indian sorghums. Heterosis in Chinese sorghums is also different from that in African sorghums and Indian sorghums. However, these data cannot be regarded as evidence of a Chinese or foreign origin for sorghum but can indicate that Chinese sorghums have high diversity and a strong evolutionary history. Except for allelic variation, there was no difference in the dominant *Tan1* sequence among wild sorghums, landraces and cultivars. This may be related to natural selection and artificial selection for tannin characteristics. *tan1-e* was only detected in 2 Chinese landraces, but *tan1-a* and *tan1-b* were not detected in Chinese landraces (Table S4 and S6). *tan1-a* and *tan1-b* may not be present in Chinese landraces. *Tan1* allelotypes were detected in 34 sterile lines and 43 restorer lines from China; 16 had the *tan1-a* allele, and 11 had the *tan1-b* allele (Table S6). Eight Chinese cultivars had the *tan1-a* allele, and 5 Chinese cultivars had the *tan1-b* allele, indicating that the *tan1-a* allele and *tan1-b* allele may have come from foreign accessions (Table S4). The *Tan1* genotypes were detected in 145 accessions of foreign sorghum (landraces and cultivars) with low tannin contents; however, there is no *tan1-e* allele in low-

tannin-producing foreign accessions (Table S4). All these results indicate that different alleles have different selective advantages in sorghum breeding and the evolutionary mode of *Tan1* is different between Chinese sorghums and foreign sorghums. *tan1-e* may have evolved in a particular way in Chinese sorghum. The two *tan1-e* landraces are from Jilin Province and Shanxi Province. The genetic background of these two accessions is quite different as there is 700 kilometers between the two provinces, although they may also have originated from a common variant ancestor. *Tan2* and six other recessive alleles (*tan2-a*, *tan2-b*, *tan2-c*, *tan2-d*, *tan2-f* and *tan2-g*) were found in the United States, West Africa, Western Europe, North America, India, China and other parts of the world. However, the recessive *tan2-e* allele was only found in Chinese landraces (Table3, Table S5). Therefore, *Tan1* and *Tan2* can be used as important clues to study the origin and evolutionary history of Chinese sorghum and foreign sorghum.

Conclusions

In our study, two new allelic variants of *Tan1* and four new allelic variants of *Tan2* were identified. Up to now, five recessive alleles of *Tan1* and seven recessive alleles of *Tan2* alleles were found, indicating that *Tan1* and *Tan2* had abundant allelic variants. Because of loss-of-function alleles in *Tan1* and *Tan2*, which lead to low or no tannin content in sorghum grains. The *tan1-e* and *tan2-e* were only found and *tan1-a* and *tan1-b* were not found in Chinese landraces, and other alleles were found in landraces or cultivars worldwide. Some *tan1*and *tan2* alleles have not been used in breeding.

Acknowledgements

This research was supported by China Agricultural Research System (CARS-06-14.5-A3), China Postdoctoral Science Foundation (2021M693847) and Central Guidance on Local Science and Technology Development Fund of Jilin Province (202002068JC).

References

Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, and Thompson WF. 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat Protoc* 1:2320-2325. 10.1038/nprot.2006.384

Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, and Lepiniec L. 2004. TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in Arabidopsis thaliana. *Plant J* 39:366-380. 10.1111/j.1365-313X.2004.02138.x

Choi J, and Kim WK. 2020. Dietary Application of Tannins as a Potential Mitigation Strategy for Current Challenges in Poultry Production: A Review. *Animals (Basel)* 10. 10.3390/ani10122389

Chung KT, Wong TY, Wei CI, Huang YW, and Lin Y. 1998. Tannins and human health: a review. *Crit Rev Food*

Sci Nutr 38:421-464. 10.1080/10408699891274273
Cos P, De Bruyne T, Hermans N, Apers S, Berghe DV, and Vlietinck AJ. 2004. Proanthocyanidins in health care: current and new trends. *Curr Med Chem* 11:1345-1359. 10.2174/0929867043365288
Dahlberg J. 2019. The Role of Sorghum in Renewables and Biofuels. *Methods Mol Biol* 1931:269-277. 10.1007/978-1-4939-9039-9_19
Dykes L. 2019. Tannin Analysis in Sorghum Grains. *Methods Mol Biol* 1931:109-120. 10.1007/978-1-4939-9039-9_8
Dykes L, and Rooney LW. 2007. Phenolic compounds in cereal grains and their health benefits. *Cereal Foods World* 52:105-111. 10.1094/CFW-52-3-0105
Gutierrez N, Avila CM, and Torres AM. 2020. The bHLH transcription factor VfTT8 underlies zt2, the locus determining zero tannin content in faba bean (Vicia faba L.). *Sci Rep* 10:14299. 10.1038/s41598-020-71070-2
Ha J, Kim M, Kim MY, Lee T, Yoon MY, Lee J, Lee YH, Kang YG, Park JS, Lee JH, and Lee SH. 2018. Transcriptomic variation in proanthocyanidin biosynthesis pathway genes in soybean (Glycine spp.). *J Sci Food Agric* 98:2138-2146. 10.1002/jsfa.8698
Habyarimana E, Dall'Agata M, De Franceschi P, and Baloch FS. 2019. Genome-wide association mapping of total antioxidant capacity, phenols, tannins, and flavonoids in a panel of Sorghum bicolor and S. bicolor x S. halepense populations using multi-locus models. *PLoS One* 14:e0225979. 10.1371/journal.pone.0225979
He F, Pan QH, Shi Y, and Duan CQ. 2008. Biosynthesis and genetic regulation of proanthocyanidins in plants. *Molecules* 13:2674-2703. 10.3390/molecules13102674
Huang Y, Wu Q, Wang S, Shi J, Dong Q, Yao P, Shi G, Xu S, Deng R, Li C, Chen H, and Zhao H. 2019. FtMYB8 from Tartary buckwheat inhibits both anthocyanin/Proanthocyanidin accumulation and marginal Trichome initiation. *BMC Plant Biol* 19:263. 10.1186/s12870-019-1876-x
Li SF, Allen PJ, Napoli RS, Browne RG, Pham H, and Parish RW. 2020. MYB-bHLH-TTG1 Regulates Arabidopsis Seed Coat Biosynthesis Pathways Directly and Indirectly via Multiple Tiers of Transcription Factors. *Plant Cell Physiol* 61:1005-1018. 10.1093/pcp/pcaa027
Morris GP, Rhodes DH, Brenton Z, Ramu P, Thayil VM, Deshpande S, Hash CT, Acharya C, Mitchell SE, Buckler ES, Yu J, and Kresovich S. 2013. Dissecting genome-wide association signals for loss-of-function phenotypes in sorghum flavonoid pigmentation traits. *G3 (Bethesda)* 3:2085-2094. 10.1534/g3.113.008417
Oliveira KG, Queiroz VA, Carlos Lde A, Cardoso Lde M, Pinheiro-Sant'Ana HM, Anunciacao PC, Menezes CB, Silva EC, and Barros F. 2017. Effect of the storage time and temperature on phenolic compounds of sorghum grain and flour. *Food Chem* 216:390-398. 10.1016/j.foodchem.2016.08.047
Rhodes DH, Hoffmann L, Jr., Rooney WL, Ramu P, Morris GP, and Kresovich S. 2014. Genome-wide association study of grain polyphenol concentrations in global sorghum [Sorghum bicolor (L.) Moench] germplasm. *J Agric Food Chem* 62:10916-10927. 10.1021/jf503651t
Schaart JG, Dubos C, Romero De La Fuente I, van Houwelingen A, de Vos RCH, Jonker HH, Xu W, Routaboul JM, Lepiniec L, and Bovy AG. 2013. Identification and characterization of MYB-bHLH-WD40 regulatory complexes controlling proanthocyanidin biosynthesis in strawberry (Fragaria x ananassa) fruits. *New Phytol* 197:454-467. 10.1111/nph.12017
Shan X, Li Y, Yang S, Gao R, Zhou L, Bao T, Han T, Wang S, Gao X, and Wang L. 2019. A functional homologue of Arabidopsis TTG1 from Freesia interacts with bHLH proteins to regulate anthocyanin and proanthocyanidin biosynthesis in both Freesia hybrida and Arabidopsis thaliana. *Plant Physiol Biochem* 141:60-72. 10.1016/j.plaphy.2019.05.015
Sun Y, Zhang D, Zheng H, Wu Y, Mei J, Ke L, Yu D, and Sun Y. 2022. Biochemical and Expression Analyses Revealed the Involvement of Proanthocyanidins and/or Their Derivatives in Fiber Pigmentation of Gossypium stocksii. *Int J Mol Sci* 23. 10.3390/ijms23021008
Wang L, Ran L, Hou Y, Tian Q, Li C, Liu R, Fan D, and Luo K. 2017. The transcription factor MYB115 contributes to the regulation of proanthocyanidin biosynthesis and enhances fungal resistance in poplar. *New Phytol* 215:351-367. 10.1111/nph.14569
Wei Z, Cheng Y, Zhou C, Li D, Gao X, Zhang S, and Chen M. 2019. Genome-Wide Identification of Direct Targets of the TTG1-bHLH-MYB Complex in Regulating Trichome Formation and Flavonoid Accumulation in Arabidopsis Thaliana. *Int J Mol Sci* 20. 10.3390/ijms20205014
Wu Y, Guo T, Mu Q, Wang J, Li X, Wu Y, Tian B, Wang ML, Bai G, Perumal R, Trick HN, Bean SR, Dweikat IM, Tuinstra MR, Morris G, Tesso TT, Yu J, and Li X. 2019. Allelochemicals targeted to balance competing selections in African agroecosystems. *Nat Plants* 5:1229-1236. 10.1038/s41477-019-0563-0
Wu Y, Li X, Xiang W, Zhu C, Lin Z, Wu Y, Li J, Pandravada S, Ridder DD, Bai G, Wang ML, Trick HN, Bean SR,

Tuinstra MR, Tesso TT, and Yu J. 2012. Presence of tannins in sorghum grains is conditioned by different natural alleles of Tannin1. *Proc Natl Acad Sci U S A* 109:10281-10286. 10.1073/pnas.1201700109

Xie P, Shi J, Tang S, Chen C, Khan A, Zhang F, Xiong Y, Li C, He W, Wang G, Lei F, Wu Y, and Xie Q. 2019. Control of Bird Feeding Behavior by Tannin1 through Modulating the Biosynthesis of Polyphenols and Fatty Acid-Derived Volatiles in Sorghum. *Mol Plant* 12:1315-1324. 10.1016/j.molp.2019.08.004

Xie Q, and Xu Z. 2019. Sustainable Agriculture: From Sweet Sorghum Planting and Ensiling to Ruminant Feeding. *Mol Plant* 12:603-606. 10.1016/j.molp.2019.04.001

Zhang L, Ding Y, Xu J, Gao X, Cao N, Li K, Feng Z, Cheng B, Zhou L, Ren M, Lu X, Bao Z, Tao Y, Xin Z, and Zou G. 2022. Selection Signatures in Chinese Sorghum Reveals Its Unique Liquor-Making Properties. *Front Plant Sci* 13:923734. 10.3389/fpls.2022.923734

Zhao ZY, Che P, Glassman K, and Albertsen M. 2019. Nutritionally Enhanced Sorghum for the Arid and Semiarid Tropical Areas of Africa. *Methods Mol Biol* 1931:197-207. 10.1007/978-1-4939-9039-9_14

Zhu F. 2019. Proanthocyanidins in cereals and pseudocereals. *Crit Rev Food Sci Nutr* 59:1521-1533. 10.1080/10408398.2017.1418284

Table 1 (on next page)

Genotype and phenotype of the 20 accessions

1 **Table 1.** Genotype and phenotype of the 20 accessions

Accessions	*Tan1*	*Tan2*	Phenotype	Origin	Germplasm type
8R156	*tan1-a*	*Tan2*	low-tannin	India	landrace
BTx623	*tan1-b*	*Tan2*	low-tannin	United States	cultivar
Tx2752	*tan1-c*	*Tan2*	low-tannin	United States	cultivar
8R111	*Tan1*	*tan2-a*	low-tannin	Senegal	landrace
8R035	*Tan1*	*tan2-a*	low-tannin	Mali	landrace
RTx430	*tan1-a*	*tan2-a*	low-tannin	United States	cultivar
8R374	*Tan1*	*Tan2*	low-tannin	China	landrace
8R336	*Tan1*	*Tan2*	low-tannin	China	landrace
JS255	*Tan1*	*Tan2*	low-tannin	China	landrace
JS257	*Tan1*	*Tan2*	low-tannin	China	landrace
JS266	*Tan1*	*Tan2*	low-tannin	China	landrace
JS273	*Tan1*	*Tan2*	low-tannin	China	landrace
8R245	*Tan1*	*Tan2*	high-tannin	China	landrace
8R249	*Tan1*	*Tan2*	high-tannin	China	landrace
8R284	*Tan1*	*Tan2*	high-tannin	China	landrace
8R243	*Tan1*	*Tan2*	high-tannin	China	landrace
8R446	*Tan1*	*Tan2*	high-tannin	China	landrace
8R312	*Tan1*	*Tan2*	high-tannin	China	landrace
SV1-5	*Tan1*	*Tan2*	high-tannin	NA	wild
TU11	*Tan1*	*Tan2*	high-tannin	NA	wild

2

Table 2 (on next page)

Genotype and phenotype of 396 accessions

1 **Table 2.** Genotype and phenotype of 396 accessions

Phenotype	*Tan1*	Accession number
low-tannin	*Tan1*	97
	tan1-a	46
	tan1-b	18
	tan1-c	14
	tan1-d	9
	tan1-e	2
high-tannin	*Tan1*	200
high-tannin (wild)	*Tan1*	10
Total		396

2

Table 3 (on next page)

Genotype and phenotype of the 72 accessions

1 **Table 3.** Genotype and phenotype of the 72 accessions

Phenotype	Genotype	Accession number
low-tannin	*Tan1/Tan2*	24
	Tan1/tan2-a	3
	Tan1/tan2-d	1
	Tan1/tan2-e	6
	Tan1/tan2-f	1
	Tan1/tan2-g	3
high-tannin	*Tan1/Tan2*	24
high-tannin (wlid)	*Tan1/Tan2*	10
Total		72

2

Figure 1

Development of molecular marker for *Tan1* and *tan1-c* in sorghum

Marker was from Wuhan Servicebio Technology Co., Ltd (GN100bp DNA Ladder I, G3365-01). A-to-T transversion at position 1054 in the coding sequence of *tan1-c* resulted in loss of a *Dde* I restriction site that was present in *Tan1*. The 165 bp PCR amplicon from *tan1-c* remained uncleaved, but the 165 bp product from *Tan1* was cleaved into 109 bp and 56 bp fragments by *Dde* I. *Tan1*: 8R191, 8R306 and TU11 (wild sorghum); *tan1-c*: Tx2752, OK11 and Tx631(Wu et al. 2019).

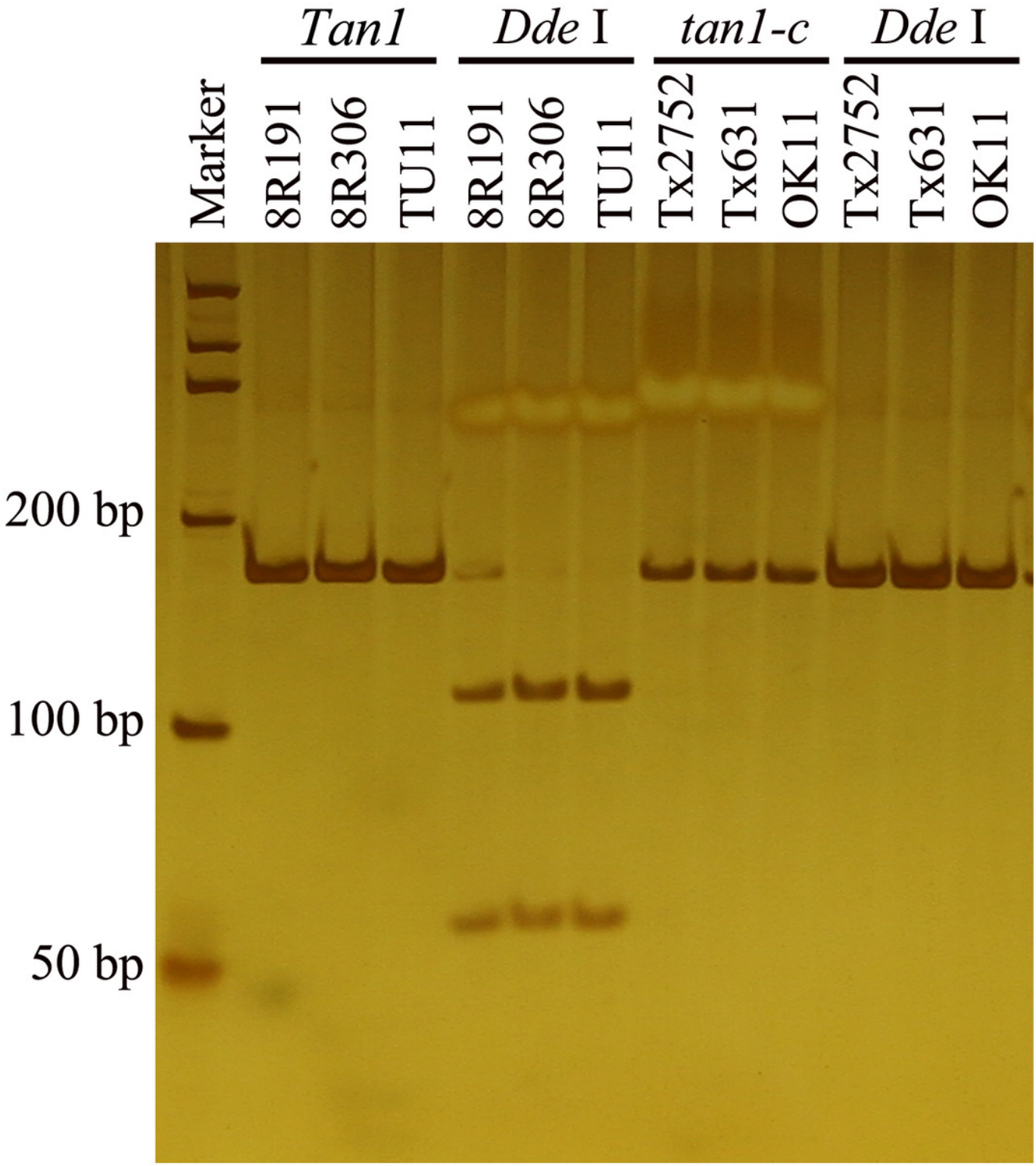
Tan1
Dde I
tan1-c
Dde I
Marker
8R191
8R306
TU11
8R191
8R306
TU11
Tx2752
Tx631
OK11
Tx2752
Tx631
OK11
200 bp
100 bp
50 bp

Figure 2

Fine mapping of the *tan1-d*

(A) ΔSNP index plot. ΔSNP index = SNP index (Tannin) − SNP index (Non-tannin). The red dashed line represents the threshold (0.49) of ΔSNP index. The area above the red dashed line is the rough mapping interval of *tan1-d* on Chr4. (B) Distribution of InDel markers in the rough mapping interval. The fine mapping interval was narrowed between InDel-7 and InDel-8. (C) Sobic.004G280700, Sobic.004G280800 (*Tan1*), Sobic.004G280900 and Sobic.004G281000 were found in the fine mapping interval.

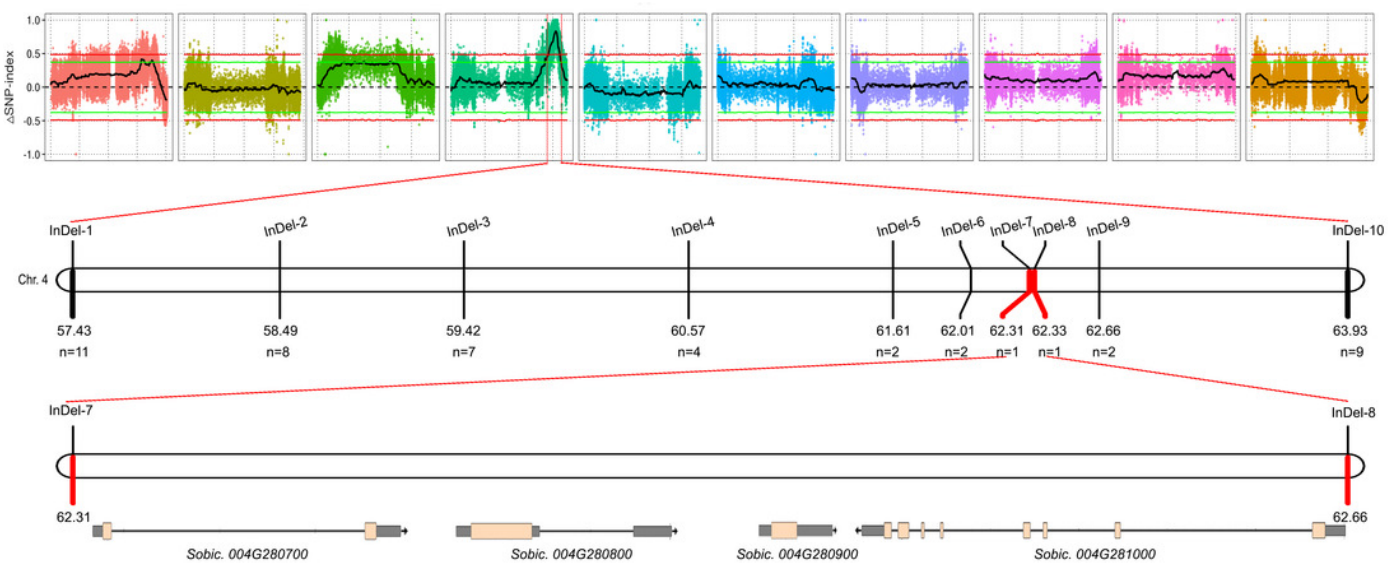

Figure 3

Comparison the coding sequences and protein sequences of Tan1, tan1-d and tan1-e

(A) Coding sequences in *Tan1*, *tan1-d* and *tan1-e*. In *tan1-d*, A-to-T(1054), GT deletion(1057 and 1058), and C-to-T(1059) were changed in the coding sequence, then TGA(1060, 1061 and 1062) stop codon frameshifted. In *tan1-e*, 10-bp(TCGACATACG) was deleted in the coding sequence between positions 771 and 780. (B) Protein and WD-40 repeat domains variation of Tan1, tan1-d and tan1-e. Compared to dominant Tan1, tan1-d had four intact WD-40 domains, but C-terminal sequence had changed greatly. The 3rd and 4th WD-40 domain, C-terminal sequence of tan1-e had altered.

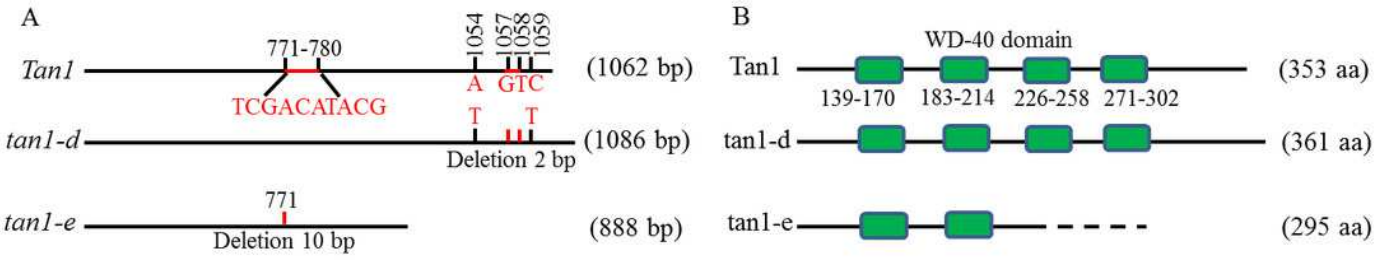

Figure 4

Variation analysis of coding sequences of *Tan1*, *tan1-c* and *tan1-d*

In *tan1-c*, A-to-T(1054), G deletion(1057), and C-to-T(1059) were changed in the coding sequence. In *tan1-d*, A-to-T(1054), GT deletion(1057 and 1058), and C-to-T(1059) were changed in the coding sequence. Sequence variation of *tan1-d* was similar to *tan1-c*, both C-terminal sequences had changed dramatically.

```
Tan1    ATGGACCTACCCAAGCCGCCGTCGACGGCCGCCTCGTCGTCGGGGGCGGAGACGCCGAACCCGCACGCCTTCACCTGCGAGCTCCCGCACTCGATCTACG 100
tan1-c  ATGGACCTACCCAAGCCGCCGTCGACGGCCGCCTCGTCGTCGGGGGCGGAGACGCCGAACCCGCACGCCTTCACCTGCGAGCTCCCGCACTCGATCTACG 100
tan1-d  ATGGACCTACCCAAGCCGCCGTCGACGGCCGCCTCGTCGTCGGGGGCGGAGACGCCGAACCCGCACGCCTTCACCTGCGAGCTCCCGCACTCGATCTACG 100

Tan1    CGCTCGCCTTCTCCCCCGGCGCGCCCGTCCTCGCCTCCGGCAGCTTCCTCGAGGACCTCCACAACCGCGTCTCCCTGCTCTCCTTCGACCCCGTCCGCCC 200
tan1-c  CGCTCGCCTTCTCCCCCGGCGCGCCCGTCCTCGCCTCCGGCAGCTTCCTCGAGGACCTCCACAACCGCGTCTCCCTGCTCTCCTTCGACCCCGTCCGCCC 200
tan1-d  CGCTCGCCTTCTCCCCCGGCGCGCCCGTCCTCGCCTCCGGCAGCTTCCTCGAGGACCTCCACAACCGCGTCTCCCTGCTCTCCTTCGACCCCGTCCGCCC 200

Tan1    CTCCGCCGCCTCCTTCCGCGCCCTCCCGGCGCTCTCCTTCGACCACCCCTACCCACCCACCAAGCTCCAGTTCAACCCGCGCGCCGCCGCGCCGTCCCTC 300
tan1-c  CTCCGCCGCCTCCTTCCGCGCCCTCCCGGCGCTCTCCTTCGACCACCCCTACCCACCCACCAAGCTCCAGTTCAACCCGCGCGCCGCCGCGCCGTCCCTC 300
tan1-d  CTCCGCCGCCTCCTTCCGCGCCCTCCCGGCGCTCTCCTTCGACCACCCCTACCCACCCACCAAGCTCCAGTTCAACCCGCGCGCCGCCGCGCCGTCCCTC 300

Tan1    CTCGCCTCCTCCGCCGACACGCTCCGCATCTGGCACGCCCCGCTCGACGACCTCTCCGCCACCGCCTCCGCGCCCGAGCTCCGCTCCGTTCTCGACAACC 400
tan1-c  CTCGCCTCCTCCGCCGACACGCTCCGCATCTGGCACGCCCCGCTCGACGACCTCTCCGCCACCGCCTCCGCGCCCGAGCTCCGCTCCGTTCTCGACAACC 400
tan1-d  CTCGCCTCCTCCGCCGACACGCTCCGCATCTGGCACGCCCCGCTCGACGACCTCTCCGCCACCGCCTCCGCGCCCGAGCTCCGCTCCGTTCTCGACAACC 400

Tan1    GCAAGGCCGCCTCCGAGTTCTGCGCGCCCCTCACCTCCTTCGATTGGAACGAGGTCGAGCCCCGCCGTATCGGGACCGCCTCCATCGACACCACCTGCAC 500
tan1-c  GCAAGGCCGCCTCCGAGTTCTGCGCGCCCCTCACCTCCTTCGATTGGAACGAGGTCGAGCCCCGCCGTATCGGGACCGCCTCCATCGACACCACCTGCAC 500
tan1-d  GCAAGGCCGCCTCCGAGTTCTGCGCGCCCCTCACCTCCTTCGATTGGAACGAGGTCGAGCCCCGCCGTATCGGGACCGCCTCCATCGACACCACCTGCAC 500

Tan1    CGTCTGGGACATCGATCTCGGCGTCGTGGAGACGCAGCTCATCGCGCACGACAAGGCCGTCCACGACATCGCCTGGGGGGAGGCCGGGGTCTTCGCCTCC 600
tan1-c  CGTCTGGGACATCGATCTCGGCGTCGTGGAGACGCAGCTCATCGCGCACGACAAGGCCGTCCACGACATCGCCTGGGGGGAGGCCGGGGTCTTCGCCTCC 600
tan1-d  CGTCTGGGACATCGATCTCGGCGTCGTGGAGACGCAGCTCATCGCGCACGACAAGGCCGTCCACGACATCGCCTGGGGGGAGGCCGGGGTCTTCGCCTCC 600

Tan1    GTGTCGGCCGACGGCTCCGTCCGCGTCTTCGACCTCCGGGACAAGGAACACTCCACCATCGTCTACGAGAGCCCCCGCCCCGACACGCCGCTCCTCAGGC 700
tan1-c  GTGTCGGCCGACGGCTCCGTCCGCGTCTTCGACCTCCGGGACAAGGAACACTCCACCATCGTCTACGAGAGCCCCCGCCCCGACACGCCGCTCCTCAGGC 700
tan1-d  GTGTCGGCCGACGGCTCCGTCCGCGTCTTCGACCTCCGGGACAAGGAACACTCCACCATCGTCTACGAGAGCCCCCGCCCCGACACGCCGCTCCTCAGGC 700

Tan1    TGGCGTGGAACCGCTCTGACCTCCGCTATATGGCCGCGCTGCTCATGGACAGCAGCGCCGTCGTCGTGCTCGACATACGTGCGCCCGGGGTGCCGGTGGC 800
tan1-c  TGGCGTGGAACCGCTCTGACCTCCGCTATATGGCCGCGCTGCTCATGGACAGCAGCGCCGTCGTCGTGCTCGACATACGTGCGCCCGGGGTGCCGGTGGC 800
tan1-d  TGGCGTGGAACCGCTCTGACCTCCGCTATATGGCCGCGCTGCTCATGGACAGCAGCGCCGTCGTCGTGCTCGACATACGTGCGCCCGGGGTGCCGGTGGC 800

Tan1    CGAGCTGCACCGGCACCGGGCGTGCGCCAACGCAGTCGCGTGGGCGCCGCAGGCCACTAGGCACCTCTGCTCGGCTGGGGACGACGGGCAGGCACTGATC 900
tan1-c  CGAGCTGCACCGGCACCGGGCGTGCGCCAACGCAGTCGCGTGGGCGCCGCAGGCCACTAGGCACCTCTGCTCGGCTGGGGACGACGGGCAGGCACTGATC 900
tan1-d  CGAGCTGCACCGGCACCGGGCGTGCGCCAACGCAGTCGCGTGGGCGCCGCAGGCCACTAGGCACCTCTGCTCGGCTGGGGACGACGGGCAGGCACTGATC 900

Tan1    TGGGAACTGCCCGAGACGGCGGCGGCTGTGCCCGCCGAGGGGATTGATCCTGTGCTAGTGTACGACGCAGGTGCCGAAATAAACCAACTTCAGTGGGCGG 1000
tan1-c  TGGGAACTGCCCGAGACGGCGGCGGCTGTGCCCGCCGAGGGGATTGATCCTGTGCTAGTGTACGACGCAGGTGCCGAAATAAACCAACTTCAGTGGGCGG 1000
tan1-d  TGGGAACTGCCCGAGACGGCGGCGGCTGTGCCCGCCGAGGGGATTGATCCTGTGCTAGTGTACGACGCAGGTGCCGAAATAAACCAACTTCAGTGGGCGG 1000

                                                 1054 1057 1058 1059
Tan1    CCGCCCACCCGGACTGGATGGCCATCGCCTTTGAGAACAAGGTCCAGCTTCTTAGGGTCTGA...................................... 1062
tan1-c  CCGCCCACCCGGACTGGATGGCCATCGCCTTTGAGAACAAGGTCCAGCTTCTTTGG.TTTGACAAGAAATTTTCTGAAGAAAGCTTGATTATCTGGAGCC 1099
tan1-d  CCGCCCACCCGGACTGGATGGCCATCGCCTTTGAGAACAAGGTCCAGCTTCTTTGG..TTGACAAGAAATTTTCTGAAGAAAGCTTGA............ 1086

Tan1    ..................................................................                                    1062
tan1-c  CTGGGACTTGGAAACCATCTGCTTGTATTCATCTTGTGTTGAGTCTGTTGACAAACCTCTTGACATAA                                   1167
tan1-d  ..................................................................                                    1086
```

Figure 5

Comparison the coding region sequences of *Tan2*, *tan2-d*, *tan2-e*, *tan2-f* and *tan2-g*

Because of synonymous mutations at position 51, 1302, 1428 and 1569 in the coding sequence, which did not affect protein function. In *tan2-d*, a 1-bp C deletion at position 563 in the coding region led to terminate prematurely. C-to-T transition at position 1366 (C AG to T AG) in the coding sequence, *tan2-e* resulted in premature termination. 8-bp (AGCTGATC) insertion between positions 1375 and 1376 in the coding region, *tan2-f* had a frameshift and an early termination. *tan2-g* had multiple subsititutions and insertions, containing 3-bp (GGC) insertion between positions 1579 and 1580, CC-to-GT at position 1600 and 1601, C-to-G at position 1603 and 9-bp (AGCTGGTGG) insertion between positions 1606 and 1607 in the coding region, C-terminal sequence had altered significantly.

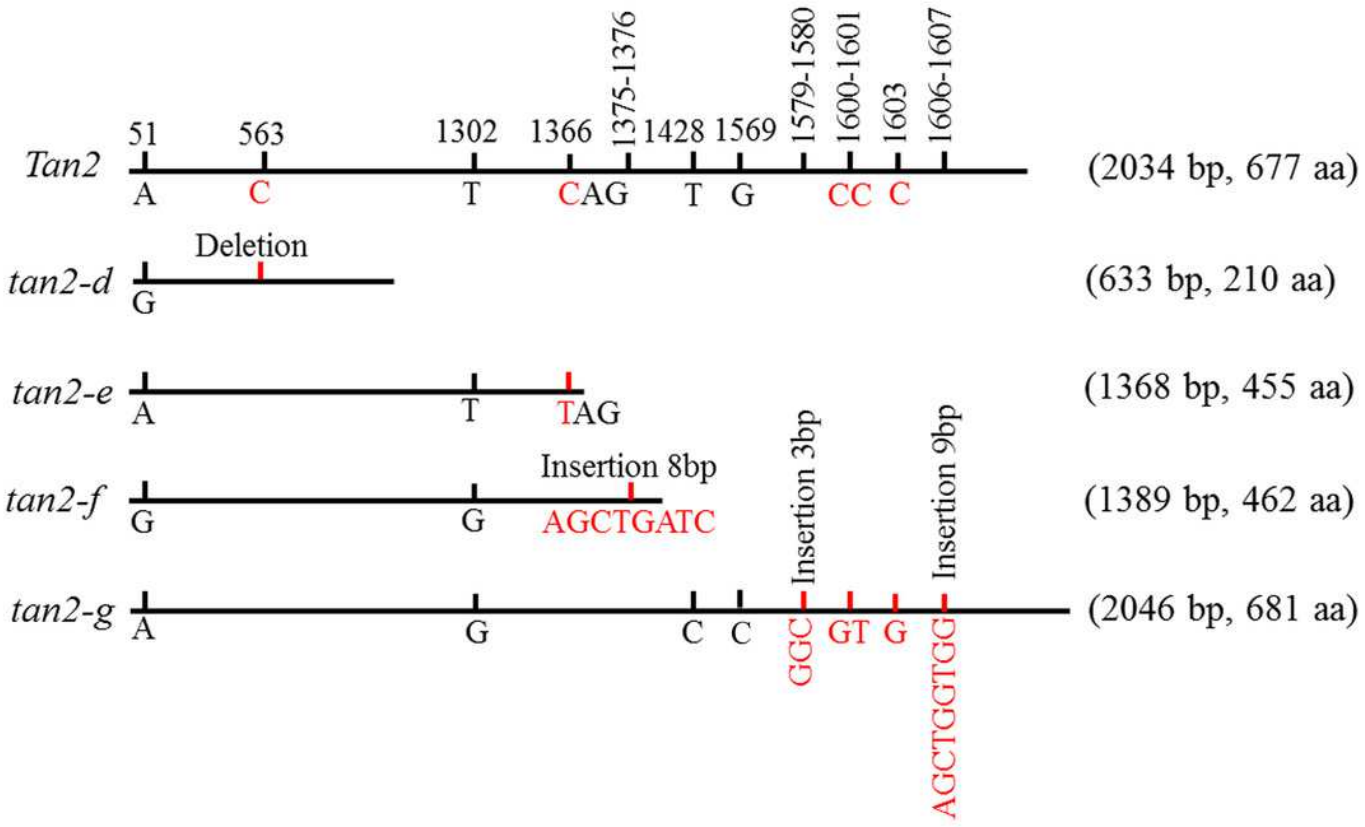

Figure 6

Comparison coding sequences of *Tan2*, *tan2-d*, *tan2-e*, *tan2-f* and *tan2-g*

Because of synonymous mutations at position 51, 1302, 1428 and 1569 in the coding sequence, which did not affect protein function. In *tan2-d*, a 1-bp C deletion at position 563 in the coding region led to terminate prematurely. C-to-T transition at position 1366 (C AG to T AG) in the coding sequence, *tan2-e* resulted in premature termination. 8-bp (AGCTGATC) insertion between positions 1375 and 1376 in the coding region, *tan2-f* had a frameshift and an early termination. *tan2-g* had multiple subsititutions and insertions, containing 3-bp (GGC) insertion between positions 1579 and 1580, CC-to-GT at position 1600 and 1601, C-to-G at position 1603 and 9-bp (AGCTGGTGG) insertion between positions 1606 and 1607 in the coding region, C-terminal sequence had altered significantly.


Tan2
tan2-d
tan2-e
tan2-f
tan2-g