Genome-wide variations analysis of special waxy sorghum cultivar Hongyingzi for brewing Moutai liquor

Can Wang, Lingbo Zhou, Xu Gao, Yanqing Ding, Bin Cheng, Guobing Zhang,
Ning Cao, Yan Xu, Mingbo Shao, Liyi Zhang*

Institute of Upland Food Crops, Guizhou Academy of Agricultural Sciences, Guiyang 550006, Guizhou, China

* lyzhang1997@hotmail.com

Funding: This research was supported by the National Natural Science Foundation of China (31660400), Special Funds for Guizhou Academy of Agricultural Sciences (QNKYYZX2014034), Science and Technology Program of Guizhou Province (QKHFQ20184005), Special Funds for the Central Government Guides Local Science and Technology Development (QKZYD20184003), and Talent Base for Germplasm Resources Utilization and Innovation of Characteristic Plant in Guizhou (RCJD2018-14).

Abstract

Hongyingzi is a special waxy sorghum (*Sorghum bicolor* L. Moench) cultivar for brewing Moutai liquor. For an overall understanding of the whole genome of Hongyingzi, we performed whole-genome resequencing technology with 56.10 X depth to reveal its comprehensive variations. Compared with the BTx623 reference genome, 2.48% of genome sequences were altered in the Hongyingzi genome. Among these alterations, there were 1885774 single nucleotide polymorphisms (SNPs), 309381 small fragments insertions and deletions (Indels), 31966 structural variations (SVs), and 217273 copy number variations (CNVs). These alterations conferred 29614 genes variations. It was also predicted that 35 genes variations were related to the multidrug and toxic efflux (MATE) transporter, chalcone synthase (CHS), ATPase isoform 10 (AHA10) transporter, dihydroflavonol-4-reductase (DFR), the laccase 15 (LAC15), flavonol 3′-hydroxylase (F3′H), flavanone 3-hydroxylase (F3H), *O*-methyltransferase (OMT), flavonoid 3′5′ hydroxylase (F3′5′H), UDP-glucose:sterol-glucosyltransferase (SGT), flavonol synthase (FLS), and chalcone isomerase (CHI) involved in the tannin synthesis. These results would provide theoretical supports for the molecular markers developments and gene function studies related to the liquor-making traits, and the genetic improvement of waxy sorghum based on the genome editing technology.

Introduction

Sorghum [*Sorghum bicolor* (L.) Moench] is the fifth largest grain crop in the world after corn, wheat, rice, and barley, which is widely distributed in the arid and semi-arid regions of the tropics, and also one of the earliest cultivated cereal crops in China [1]. It has become a model crop for genome research of cereal crops because of its wide adaptability to environment, strong stress resistance, rich resources, and relatively small genome [2, 3]. According to different purposes, sorghum are generally divided into three types, namely sweet sorghum, feed sorghum, and grain sorghum. In grain sorghum, cultivars with amylose content between 0% and 5% are called waxy sorghum [4]. Waxy sorghum is one of the main raw materials for Moutai-flavor liquor and Luzhou-flavor liquor production due to its high amylopectin and tannin contents [5, 6]. In recent years, the undiversified main liquor-making waxy sorghum cultivar and its continuous degradation phenomenon has affected the supply of raw materials for liquor-making waxy sorghum and restricted the development of liquor enterprises [7]. Therefore, investigation of waxy sorghum genetic resources is a crucial measure for better straight evolution, genetic studies, and liquor-making waxy sorghum breeding strategies.

Genetic variation is a kind of variation that can be passed on to offspring due to the changes of genetic material in organisms and leads to the genetic diversity at different levels. There are many types of genetic variation in the genome, from microscopic chromosome inversion to single nucleotide mutation. With the development of genomics, the information of genetic variation that can be studied has become more comprehensive, such as single nucleotide

polymorphism (SNP), small fragments insertion and deletion (Indel), structural variation (SV), and copy number variation (CNV) [8-10]. SNP is a kind of DNA sequence polymorphism caused by single base conversion or transversion, which is a new generation of molecular marker after restriction fragment length polymorphism (RFLP) and simple sequence repeats (SSR). It has been widely used in the construction of genetic linkage map, quantitative trait locus (QTL) mapping, genome-wide association study (GWAS), population genetic structure study, and genetic diversity analysis due to its characteristics of easy detection, large quantity, rich polymorphism, large flux, and wide distribution in genome [11-13]. Indel is a molecular biology term for an insertion or deletion of nucleotide fragments of different sizes at the same site in the genome sequence between the same or closely related species, which is widely distributed across the genome and occurs in a high density and large numbers in a genome. It has been applied to genetic analyses of animal and plant populations, molecular assisted crops and farmed animal breeding, human forensic genetics, and medical diagnostics because of its abundance, convenient typing platform, high accuracy, and good stability [14-16]. SV is operationally defined as genomic alterations that involve segments of DNA that are larger than 1 kb, and can be microscopic or submicroscopic, which mainly includes inversion, insertion, deletion, duplication, and other gene rearrangement. It can produce new genes, alter gene dosage and structure, and regulate gene expression elements, and have a significant impact on phenotypic variation and gene expression [17, 18]. CNV is a kind of genomic structural variation originated from gain or loss of DNA segments larger than 1 kb caused by genomic rearrangement, which has been reported to be associated with human complex diseases and widely used for prevention and clinical diagnoses of human diseases since it was first discovered in human populations. It is also widely found in the plant genomes, such as *Arabidopsis*, rice, corn, soybean, wheat, and cucumber, and its own gained or lost copies may result in the alteration of gene dosage and abundance of its transcript, and thus lead to the significant phenotypic variation of height, flowering time, and dormancy in plants [19, 20]. With the rapid development of molecular biology, whole-genome resequencing technology has been applied to genome-wide variations analysis in *Arabidopsis*, rice, maize, tomato, and other plants [8, 21-23]. The whole genome sequences of grain sorghum cultivar BTx623 has provided a template for genome-wide variations analysis in sorghum [24], and the first genome-wide variations analysis of sorghum was reported by [25].

Hongyingzi, a special waxy sorghum cultivar for brewing Moutai liquor containing 83.40% total starch, 80.29% amylopectin/total starch ratio, and 1.61% tannin. The genome-wide variation of Hongyingzi is not fully understood, yet it is necessary for liquor-making waxy sorghum functional genomic research and breeding. Here, we used whole-genome resequencing technology to study the whole genome variation of Hongyingzi, and discovered potential genome regions and metabolic pathways associated with liquor-making traits.

Materials and methods

Plant materials and whole-genome resequencing

Two sorghum cultivars were used in this study. Hongyingzi, approved by the Guizhou Crop Cultivar Approval Committee (Guiyang, Guizhou Province, China) in 2008, is a medium maturity waxy sorghum cultivar special used for brewing Moutai liquor and developed by Renhuai Fengyuan Organic Sorghum Breeding Center at Guizhou, China in 2008 [26]. BTx623 is an excellent grain sorghum cultivar used for whole-genome sequencing by the Joint Genome Institute and for constructing several mapping populations [25, 27].

Sees of Hongyingzi were sterilized by soaking in 0.1% mercury dichloride for 15 min, and then rinsed with distilled water for ten times. Next, seeds were placed in a germination box lined with three layers of filter paper and added 15 mL distilled water. The germination box was placed in the RXZ-1000B artificial climate box for cultivating 10 days as following parameters settings, day/night temperature is 28°C/25°C, light/dark time is 12 h/12 h, humidity is 85%, and light intensity is 340 μmol m^{-2} s^{-1}. The 10-day-old healthy seedlings were harvested for DNA extraction using the CTAB (Hexadecyl trimethyl ammonium bromide) buffer method [28]. The DNA purity was determined by 0.8% agarose gel 100 V electrophoresis for 40 min and DNA concentration was determined by Qubit® 2.0 fluorescent meter (Invitrogen, Carlsbad, USA). Following quality assessment, the genomic DNA was

randomly broken into 350 bp fragments by Covaris ultrasonic crushing apparatus and DNA fragments were end repaired, added ployA tail, added sequencing connector, purification, and PCR amplification to complete the establishment of the library. The constructed library was used to paired-end PE150 sequencing on Illumina HiSeq 4000 sequencing platform. The BTx623 reference genome sequences were downloaded from the https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_SbicolorRio_er.

Genomic variation detection and annotation

Bioinformatics analysis was carried out by Beijing Novogene technology co., LTD (Beijing, China). The original image data generated by the sequencing machine were converted into sequence data via base calling (Illumina pipeline CASAVA v1.8.2) and then subjected to quality control (QC) procedure to remove unusable reads according to following criteria: the reads contain the Illumina library construction adapters, the reads contain more than 10% unknown bases (N bases), and one end of the read contain more than 50% of low quality bases (sequencing quality value ≤ 5). After filtration, sequencing reads were aligned to the BTx623 reference genome using BWA [29] with default parameters. Subsequent processing, including duplicate removal was proformed using SAMtools [30] and PICARD (http://picard.sourceforge.net). The raw SNP/Indel sets were called by SAMtools with the parameters as '-q1 -C50 -m2 -F0.002 -d1000', and then filtered this sets using the following criteria: the mapping quality > 20 and the depth of the variate position > 4. BreakDancer [31] and CNVnator [32] were used for SV and CNV detections respectively. ANNOVAR [33] was used for functional annotation of variants. The UCSC known genes were used for gene and region annotations.

Gene variation analysis

Using the BTx623 gene set as the reference, genes with non-synonymous SNPs and Indels in coding regions identified in the Hongyingzi were selected as the candidate gene set. These genes were then aligned to the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database using Blast for clustering analysis [34, 35].

Results

Genome-wide identification of genetic variations in Hongyingzi

The whole genome of Hongyingzi was resequenced using Illumina Genome Analyser sequencing technology. The genome size of the BTx623 reference genome is 732152042 bp. Resequencing yielded 45.84 Gb of raw data, which comprised 45.79 Gb of high quality clean data (Table 1). There was a high sequencing quality (Q20 ≥ 97.55%, Q30 ≥ 93.10) and the GC content was 44.30%. The results showed that 97.52% of the Hongyingzi genome sequences (297504853 mapped reads) were identical to BTx623, average depth 56.10 X, with 95.94% of coverage at 1 X and 94.17% of coverage at least 4 X (Table 2). With these reads and the information from the BTx623 reference genome, large quantities of SNPs, Indels, SVs, and CNVs were identified (Fig. 1). Compared with the BTx623 reference genome, we finally found 1885774 SNPs, 309381 Indels, 31966 SVs, and 217273 CNVs in Hongyingzi.

SNPs in the Hongyingzi genome

A total of 1885774 SNPs were identified in the Hongyingzi genome, including 1230508 transitions and 655266 transversions (Fig. 2A). Besides, there were 1401089 homozygous SNPs and 484685 heterozygous SNPs (Fig. 2B), and the het rate was 0.066%. As shown by annotations of SNPs detected in Hongyingzi (Table 3), there were 1515993 SNPs mutation in intergenic, 89326 SNPs in 1 kb of upstream, 75170 SNPs in 1 kb of downstream, and 6344 SNPs mutated in both 1 kb of upstream and downstream. We found that 76528 SNPs were mutated in exonic, including 38176 synonymous SNPs, 37774 non-synonymous SNPs, 453 SNPs related to gain of stop codons, and 125 SNPs related to loss of stop codons. We also found that there were 122211 SNPs mutation in intronic and 202 SNPs in splicing sites. Besides, the proportion of C:G>T:A type was observed to be the highest (Fig. 3).

Indels in the Hongyingzi genome

A total of 309381 Indels containing 149071 insertions and 160310 deletions, was uncovered in the Hongyingzi genome (Fig. 4A). These Indels also included 309361 homozygous and 20 heterozygous Indels (Fig. 4B), and the het rate was 0.0065%. Annotation analysis (Table 4) showed that there were 190165, 38198, 28361, and 2779 Indels mutated in intergenic, 1 kb of upstream, 1 kb of downstream, and both 1 kb of upstream and downstream, respectively. We found that 9375 Indels were mutated in exonic, in which 103 Indels were related to gain of stop codons, 22 Indels were related to loss of stop codons, 1354 insertions and 1476 deletions might lead to frameshift, and 3219 insertions and 3201 deletions might lead to non-frameshift. We also found 40223 Indels were mutated in intronic and 189 Indels did in splicing sites. Besides, the proportion of 1 bp (Fig. 5A) and 3 bp (Fig. 5B) Indels were observed to be the highest in whole genome and coding regions, respectively.

SVs in the Hongyingzi genome

A total of 31966 SVs were identified in the Hongyingzi genome, including 70 insertions, 15975 deletions, 1948 inversions, 4938 intrachromosomal translocations, and 9035 interchromosomal translocations (Fig. 6). As shown by annotations of SVs detected in Hongyingzi (Table 5), there were 9661 SVs mutation in intergenic, 1915 in 1 kb of upstream, 1460 in 1 kb of downstream, and 176 in both 1 kb of upstream and downstream. We also found that there were 3657 SVs mutation in exonic, 1119 in intronic, and 5 in splicing sites.

CNVs in the Hongyingzi genome

A total of 217273 CNVs including 4966 duplications and 16307 deletions was uncovered in the Hongyingzi genome (Fig. 7). Annotation analysis (Table 6) showed that there were 17082, 985, 789, and 96 CNVs mutated in intergenic, 1 kb of upstream, 1 kb of downstream, and both 1 kb of upstream and downstream, respectively. We also found that there were 1822 CNVs and 496 CNVs mutated in exonic and intronic.

Functional clustering of gene variations

Compared to the BTx623 reference genome, 29614 genes variations were identified in the Hongyingzi genome (Table 7). Of which, 14028, 25166 and 3948 was caused by SNPs, Indels, and 3948 SVs, respectively. GO annotation showed that SNPs and Indels were distributed among different gene ontologies (Fig. 8). In cellular component ontology, the cell and cell part contained the majority of gene variations with 19.06% SNPs and 23.01% Indels. Extracellular matrix contained a lower rate of variation. In molecular function ontology, binding and catalytic activity had a higher rate of variation. Binding included 40.77% and 37.56% of variation in SNPs and Indels, while catalytic activity did 34.01% 31.67% of variation in SNPs and Indels. In biological process ontology, metabolic process and cellular process had a high rate of variation. Metabolic process term included 39.00% and 36.07% of variation in SNPs and Indels, while cellular process did 39.05% and 35.99% of variation in SNPs and Indels. In KEGG annotation, 141 gene variations caused by SNPs (Fig. 9A) involved in the ubiquitin mediated proteolysis, while 1756 caused by Indels (Fig. 9B) involved in the metabolic pathways. These variations may affect the distinguishing traits between Hongyingzi and BTx623.

Genes variations involved in tannin synthesis

Compared to the BTx623 reference genome, we found that 35 genes variations were related to the tannin synthesis in the Hongyingzi genome (Table 8). Of which, 7 genes did in the multidrug and toxic efflux (MATE) transporter, 7 involved in the chalcone synthase (CHS), 4 did in the ATPase isoform 10 (AHA10) transporter, 4 did in the dihydroflavonol-4-reductase (DFR), 3 did in the laccase 15 (LAC15), 2 did in the flavonol 3′-hydroxylase (F3′H), 2 did in the flavanone 3-hydroxylase (F3H), 2 did in the *O*-methyltransferase (OMT), 1 did in the flavonoid 3′5′ hydroxylase (F3′5′H), 1 did in the UDP-glucose:sterol-glucosyltransferase (SGT), 1 did in the flavonol synthase

(FLS), and 1 did in the chalcone isomerase (CHI).

Discussion

The rapid development of high-throughput sequencing technologies and bioinformatic tools makes it possible to understand the genetic variation and diversity of sorghum at the whole genome level, which plays an important role in enriching sorghum germplasm resources [24, 25, 36]. In this study, we used whole-genome resequencing technology to analyze the genetic variation in Hongyingzi, which is a special waxy sorghum cultivar for brewing Moutai liquor. The results showed that found that 2.48% of genome sequences were different between Hongyingzi and BTx623, and more than two million SNPs and Indels, along with large numbers of SVs and CNVs were identified. This is the first report on the genome-wide variations analysis in liquor-making waxy sorghum, which will be valuable for further genotype-phenotype studies and for molecular marker assisted breeding of liquor-making waxy sorghum.

In this study, the proportion of SNPs in intronic regions was 6.48%, which was higher than that in *Arabidopsis* [37]. Because the average intron size of sorghum is 444 bp, while the *Arabidopsis* is 168 bp [25]. A large number of SNPs was identified to alter in 202 splicing sites, 453 gain of stop codons, and 125 loss of stop codons. These alterations could lead to open reading frames extension, functional gene expression failure, or intron size increase [8, 21, 38]. Besides, the proportion of 3 bp Indels was observed to be the highest in coding regions. This might be due to the loss or increase of three bases results in the deletion or addition of a single amino acid without disrupting the overall reading frame [39], which could be a protection means to avoid the drastic changes of the genetic coding information, and then reduce damage to organisms due to natural variation. In addition, Indels with no multiples of 3 bp were rare in coding regions but relatively common in non-coding regions, because most of frameshift mutations is harmful to sorghum survival [25]. Compared to the BTx623 reference genome, a large number of SVs and CNVs was presented in the Hongyingzi genome, and the annotations of SVs and CNVs were similar to that of SNPs and Indels.

Compared to the BTx623 reference genome, there were 29614 genes variations in the Hongyingzi genome and Indels accounted for most of the genes variations. However, previous studies reported that SNPs accounted for most of the genes variations in *Arabidopsis* [40] and sorghum [25]. There are two possible reasons: 1) different materials used in different research, 2) limitations of early sequencing technology. Studies of SVs and CNVs in sorghum lag behind those in other plants. Recent studies in maize showed it potentially contributed to the heterosis during domestication and disease responses [41, 42]. Thus, we should focused on non-synonymous SNPs and Indels in coding regions for subsequent analysis of mutative genes. In our study, GO annotation showed that the mutative genes were equal distribution in different GO term. This indicates that SNPs and Indels may share similar survival and distribution patterns, although the origins and scales may different for affected genome segments.

Tannin, also known as condensed tannin or proanthocyanidins, is oligomers and polymers of flavan-3-ols [43, 44]. Sorghum has been the raw material for making famous liquor because of its grains containing tannin, and contributed special taste to Moutai-flavor liquor [45, 46]. Previous studies have mapped some gene loci associated with tannin content of sorghum. The *Tan1* gene (*Sb04g031730*) was cloned, which code a WD40 protein and control the tannin biosynthesis [43]. Two gene loci linked to tannin content were found [47]. One was named as *Sb01g001230*, coding glutathione-S-transferase, another was named as *Sb02g006390*, coding bHLH transcription factor and was isotopic with gene B_2 for color seed coat. Compared to the BTx623 reference genome, 35 genes variations were related to the tannin synthesis in the Hongyingzi genome. The genes involved in the MATE transporter, CHS, AHA10 transporter, DFR, LAC15, F3′H, F3H, OMT, F3′5′H, SGT, FLS, and CHI. Its variations would provide theoretical supports for the molecular markers developments and gene cloning, and the genetic improvement of waxy sorghum based on the genome editing technology.

Conclusions

231 This is a first report of genome-wide variations analysis in liquor-making waxy sorghum. High-density SNP, Indel,
232 SV, and CNV markers reported here will be a valuable resource for future gene-phenotype studies and the molecular
233 breeding of liquor-making waxy sorghum. Genes variations involved in tannin synthesis reported here will provide
234 theoretical basis for marker developing and gene cloning.
235
236 ## References

237 1. Zou GH, Zhai GW, Feng Q, Yan S, Wang AH, Zhao Q, et al. Identification of QTLs for eight agronomically important traits using
238 an ultra-high-density map based on SNPs generated from high-throughput sequencing in sorghum under contrasting photoperiods.
239 J Exp Bot. 2012; 63(15):5451–5462.
240 2. Nagaraja Reddy R, Madhusudhana R, Murali Mohan S, Chakravarthi DVN, Mehtre SP, Seetharama N, et al. Mapping QTL for grain
241 yield and other agronomic traits in post-rainy sorghum [*Sorghum bicolor* (L.) Moench]. Theor Appl Genet 2013; 126(8):1921–1939.
242 3. Boyles RE, Pfeiffer BK, Cooper EA, Rauh BL, Zielinski KJ, Myers MT, et al. Genetic dissection of sorghum grain quality traits
243 using diverse and segregating populations. Theor Appl Genet. 2017; 130(4):697–716.
244 4. Wang C, Zhou LB, Zhang GB, Xu Y, Zhang LY, Gao X, et al. Drought resistance identification and drought resistance indices
245 screening of liquor-making waxy sorghum resources at adult plant stage. Sci Agric Sin. 2017; 50(8):1388–1402.
246 5. Ni XL, Zhao GL, Liu TP, Long WJ, Hu JL, Ding GX. Genetic diversity analysis of glutinous sorghum germplasm resources based
247 on SSR markers. Agric Sci Technol. 2016; 17(3):499–504.
248 6. Wang C, Zhou LB, Zhang GB, Xu Y, Zhang LY, Gao X, et al. Optimal fertilization for high yield and good quality of waxy sorghum
249 (*Sorghum bicolor* L. Moench). Field Crops Res. 2017; 203:1–7.
250 7. Gao X, Zhou LB, Zhang GB, Shao MB, Zhang LY. Genetic diversity and population structure of grain sorghum germplasm resources
251 based on SSR marker. Guizhou Agric Sci. 2016; 44(9):13–19.
252 8. Cheng ZX, Lin JC, Lin TX, Xu M, Huang ZW, Yang ZJ, et al. Genome-wide analysis of radiation-induced mutations in rice (*Oryza*
253 *sativa* L. ssp. indica). Mol Biosyst. 2014; 10(4):795–805.
254 9. Vergara IA, Tarailo-Graovac M, Frech C, Wang J, Qin ZZ, Zhang T, et al. 2014. Genome-wide variations in a natural isolate of the
255 nematode *Caenorhabditis elegans*. BMC Genomics. 2014; 15:255.
256 10. Shao XH, Hu CH, Sheng O, Bi FC, Deng GM, Yang QS, et al. 2018. Genome-wide variations of triploid banana (AAA group)
257 'Grand Nain' by whole-genome resequencing. Plant Physiol J. 2018; 54(4):581–593.
258 11. Liu CG, Zhang GQ. Single nucleotide polymorphism (SNP) and its application in rice. Hereditas. 2006; 28(6):737–744.
259 12. Liu LZ, Qu CM, Wittkop B, Yi B, Xiao Y, He YJ, et al. A high-density SNP map for accurate mapping of seed fibre QTL in
260 *Brassica napus* L. PLoS ONE. 2013; 8(12):e83052.
261 13. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, et al. Population genomic and genome-wide association
262 studies of agroclimatic traits in sorghum. Proc Natl Acad Sci USA. 2013; 110(2):453–458.
263 14. Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL. *Arabidopsis* map-based cloning in the post-genome era. Plant
264 Physiol. 2002; 129(2):440–450.
265 15. Weber JL, David D, Heil J, Fan Y, Zhao CF, Marth G. Human diallelic insertion/deletion polymorphisms. Am J Hum Genet. 2002;
266 71(4):854–862.
267 16. Yang J, He J, Wang DB, Shi E, Yang WY, Geng QF, et al. Progress in research and application of InDel markers. Biodivers Sci.
268 2016; 24(2):237–243.
269 17. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006; 7:85–97.
270 18. He YS, Zhang W, Yang ZQ. Structural variation in the human genome. Hereditas. 2009; 31(8):771–778.
271 19. Jiao YP, Zhao HN, Ren LH, Song WB, Zeng B, Guo JJ, et al. Genome-wide genetic changes during modern breeding of maize.
272 Nat Genet. 2012; 44:812–815.
273 20. Yang HJ, Zhang DQ. Copy number variations in plant genomes. Mol Plant Breeding 2015; 13(8):1895–1910.
274 21. Lai JS, Li RQ, Xu X, Jin WW, Xu ML, Zhao HN, et al. Genome-wide patterns of genetic variation among elite maize inbred lines.
275 Nat Genet. 2010; 42:1027–1030.
276 22. Long Q, Rabanal FA, Meng DZ, Huber CD, Farlow A, Platzer A, et al. 2013. Massive genomic variation and strong selection in

Arabidopsis thaliana lines from Sweden. Nat Genet. 2013; 45:884–890.

23. Ercolano MR, Sacco A, Ferriello F, D′Alessandro R, Tononi P, Traini A, et al. 2014. Patchwork sequencing of tomato San Marzano and Vesuviano varieties highlights genome-wide variations. BMC Genomics. 2014; 15:138.
24. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009; 457:551–556.
25. Zheng LY, Guo XS, He B, Sun LJ, Peng Y, Dong SS, et al. Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). Genome Biol. 2011; 12:R114.
26. Huang WP, Chen Q. High yield cultivation technique of new sorghum cultivar Hongyingzi. Agric Technol Serv. 2010; 27(4):427, 442
27. Paterson AH. 2008. Genomics of sorghum. Int J Plant Genomics. 2008; 2008:362451.
28. Zhang LG, Cheng ZJ, Qin RZ, Qiu Y, Wang JL, Cui XK, et al. Identification and characterization of an epi-allele of *FIE1* reveals a regulatory linkage between two epigenetic marks in rice. Plant Cell. 2012; 24(11):4407–4421.
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–1760.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N. et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16):2078–2079.
31. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009; 6:677–681.
32. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011; 21:974–984.
33. Wang K, Li MY, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164.
34. Mao XZ, Cai T, Olyarchuk JG, Wei LP. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics. 2005; 21(19):3787–3793.
35. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genom. 2008; 2008:619832.
36. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015; 58(4):586–597.
37. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science. 2007; 317(5836):338–342.
38. Wang L, Hao LX, Li X, Hu SL, Ge SX, Yu JK. SNP deserts of Asian cultivated rice: genomic regions under domestication. J Evol Biol. 2009; 22(4):751–761.
39. Hu M, Yao SL, Cheng XH, Liu YY, Ma LX, Xiang Y, et al. Genomic variation of spring，semi-winter and winter *Brassica napus* by high-depth DNA re-sequencing. Chin J Oil Crop Sci. 2018; 40(4):469–478.
40. Chang FQ, Liu XM, Li YX, Jia GX, Ma J, Liu S, et al. Analysis of Low energy N^+ irradiation induced genome DNA variation of *Arabidopsis thaliana*. Sci China Ser C. 2003; 33(2):117–124.
41. Springer NM, Ying K, Fu Y, Ji TM, Yeh CT, Jia Y, et al. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 2009; 5(11):e1000734.
42. Beló A, Beatty MK, Hondred D, Fengler KA, Li B, Rafalski A. Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet. 2010; 120(2):355.
43. Wu YY, Li XR, Xiang WW, Zhu CS, Lin ZW, Wu Y, et al. Presence of tannins in sorghum grains is conditioned by different natural alleles of *Tannin1*. Proc Natl Acad Sci USA. 2012; 109(26):10281–10286.
44. Zhang CL, Li YF, Zhao WJ, Zhao L, Wang C, Liang D, et al. Molecular genetic basis for biotechnological improvement of grain quality characteristics in sorghum. Plant Physiol J. 2015; 51(5):610–616.
45. Xiong XQ, Chen RX, Yang F, Liu ZS, Zhou YF. Inspection and identification of germplasm resources of wine-making sorghum in Guizhou. J Mt Agric Biol. 2003; 22(2):117–121.
46. Bai CM, Wang CY, Wang P, Zhu ZX, Lu XC. QTLs analysis of tannin content and color of grain in sorghum. J Plant Genet Res.

2017; 18(5):860–866.

47. Morris GP, Rhodes DH, Brenton Z, Ramu P, Thayil VM, Deshpande S, et al. Dissecting genome-wide association signals for loss-of-function phenotypes in sorghum flavonoid pigmentation traits. Genes Genom Genet. 2013; 3(11):2085–2094.

Table 1

Summary of resequencing data of Hongyingzi.

Raw base (Gb)	Clean base (Gb)	Effect rete (%)	Error rate (%)	Q20 (%)	Q30 (%)	GC content (%)
45.84	45.79	99.82	0.03	97.55	93.10	44.30

Table 2

Sequence alignment of Hongyingzi to BTx623.

Mapped reads	Total reads	Mapping rate (%)	Average depth (X)	Coverage at least 1 X (%)	Coverage at least 4 X (%)
297504853	305064750	97.52	56.10	95.94	94.17

Table 3

Annotations of SNPs detected in Hongyingzi.

Category	Numbers of SNPs	Region
Intergenic	1515993	
1 kb of upstream	89326	
1 kb of downstream	75170	
Both 1 kb of upstream and downstream	6344	
Gain of stop codons	453	Coding regions
Loss of stop codons	125	Coding regions
Synonymous	38176	Coding regions
Non-synonymous	37774	Coding regions
Intronic	122211	
Splicing sites	202	

Table 4

Annotations of Indels detected in Hongyingzi.

Category	Numbers of Indels	Region
Intergenic	190165	
1 kb of upstream	38198	
1 kb of downstream	28361	
Both 1 kb of upstream and downstream	2779	
Gain of stop codons	103	Coding regions
Loss of stop codons	22	Coding regions
Frameshift (insertions)	1354	Coding regions
Frameshift (deletions)	1476	Coding regions
Non-frameshift (insertions)	3219	Coding regions
Non-frameshift (deletions)	3201	Coding regions
Intronic	40223	
Splicing sites	189	

525

526 **Table 5**

527 Annotations of SVs detected in Hongyingzi.

Category	Numbers of SVs
Intergenic	9661
1 kb of upstream	1915
1 kb of downstream	1460
Both 1 kb of upstream and downstream	176
Exonic	3657
Intronic	1119
Splicing sites	5

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

Table 6

Annotations of CNVs detected in Hongyingzi.

Category	Numbers of CNVs
Intergenic	17082
1 kb of upstream	985
1 kb of downstream	789
Both 1 kb of upstream and downstream	96
Exonic	1822
Intronic	496

602

603 **Table 7**

604 Summary of gene variations in Hongyingzi.

Variation types			Total
SNPs	Indels	SVs	
14028	25166	3948	29614

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

Table 8

Summary of tannin synthesis related gene variations.

Gene name	Chromosome	Annotation	Variation type	Variation information
Sobic.001G012600	1	*SbMATE*	Non-synonymous SNP	1175307 bp, G/A
Sobic.001G185400	1	*SbMATE*	Non-synonymous SNP	15851598 bp, G/A; 15851639 bp, C/T; 15851643 bp, G/C; 15851644 bp, G/C; 15857015 bp, G/A
			Non-frameshift insertion	15851633 bp, -/GGT
Sobic.001G185500	1	*SbMATE*	Non-frameshift insertion	15867230 bp, -/GCACGG
Sobic.001G185600	1	*SbMATE*	Non-synonymous SNP	15877514 bp, G/T; 15877530 bp, T/A; 15879898 bp, T/G
			Non-frameshift deletion	15880626 bp, ACCGGCGCC/-
Sobic.004G349550	4	*SbMATE*	Non-frameshift insertion	67834717 bp, -/GCTGCT
Sobic.004G349600	4	*SbMATE*	Non-synonymous SNP	67848132 bp, C/G; 67848135 bp, C/G; 67848250 bp, T/G
Sobic.007G165500	7	*SbMATE*	Non-synonymous SNP	60025990 bp, C/A
Sobic.001G360800	1	*SbF3'5'H*	Non-synonymous SNP	65069116 bp, T/C
Sobic.001G543900	1	*SbAHA10*	Non-synonymous SNP	80740375 bp, G/C
Sobic.003G436400	3	*SbAHA10*	Non-frameshift insertion	73733903 bp, -/CCG
Sobic.010G063700	10	*SbAHA10*	Non-synonymous SNP	5033142 bp, C/A; 5033614 bp, C/T; 5033877 bp, C/A; 5034044 bp, C/T; 5034053 bp, G/C
			Frameshift insertion	5032581 bp, -/GC; 5032758 bp, -/GAGC; 5033017 bp, -/ATCT
			Non-frameshift deletion	5032669 bp, GTGCTGTTC/-
			Non-frameshift insertion	5033742 bp, -/GGG; 5034105 bp, -/TTCCAC
			Gain of stop codons	5034223 bp, -/CTATTTCA
Sobic.010G207800	10	*SbAHA10*	Non-synonymous SNP	55088876 bp, T/C
Sobic.002G117500	2	*SbSGT*	Non-synonymous SNP	14508960 bp, C/A
Sobic.002G310500	2	*SbCHS*	Non-synonymous SNP	68442264 bp, A/C; 68442283 bp, G/A
Sobic.004G179000	4	*SbCHS*	Non-synonymous SNP	53190344 bp, T/C
Sobic.005G135600	5	*SbCHS*	Non-synonymous SNP	58503342 bp, T/A; 58503472 bp, T/C; 58503507 bp, G/A; 58503555 bp, C/G
Sobic.005G136200	5	*SbCHS*	Non-synonymous SNP	58859286 bp, C/G
Sobic.005G136300	5	*SbCHS*	Non-synonymous SNP	58881162 bp, G/A
Sobic.005G137100	5	*SbCHS*	Non-synonymous SNP	58943632 bp, C/T
Sobic.008G036800	8	*SbCHS*	Non-synonymous SNP	3477776 bp, G/A
			Non-frameshift deletion	3477795 bp, ACG/-
Sobic.003G230900	3	*SbDFR*	Non-synonymous SNP	57029960 bp, C/T
Sobic.003G231000	3	*SbDFR*	Non-frameshift deletion	57041941 bp, CTGGGA/-
Sobic.004G050200	4	*SbDFR*	Non-frameshift deletion	4052019 bp, AAC/-
Sobic.009G043800	9	*SbDFR*	Non-synonymous SNP	4149752 bp, T/C; 4149842 bp, G/A; 4149896 bp, G/T; 4149998 bp, T/C; 4150031 bp, C/G
Sobic.004G200900	4	*SbF3'H*	Non-synonymous SNP	55234140 bp, T/G
			Non-frameshift deletion	55233739 bp, CGGGAA/-
Sobic.009G162500	9	*SbF3'H*	Non-synonymous SNP	51944205 bp, A/G; 51948174 bp, C/G
Sobic.004G236000	4	*SbLAC15*	Non-synonymous SNP	58382355 bp, G/A; 58382419 bp, G/A; 58383602 bp, A/G; 28383682 bp, G/T
Sobic.004G236100	4	*SbLAC15*	Non-synonymous SNP	58391947 bp, C/T
			Frameshift deletion	58392294 bp, CTAC/-
Sobic.005G156700	5	*SbLAC15*	Non-synonymous SNP	62814031 bp, G/A; 62814043 bp, A/G; 62814250 bp. C/G
			Non-frameshift deletion	62816156 bp, CGTCAACGT/-
			Frameshift deletion	62813716 bp, C/-; 62813926 bp, A/-; 62814183 bp, C/-
			Frameshift insertion	62813832 bp, -/A; 62814474 bp, -/TA
Sobic.004G310100	4	*SbFLS*	Non-synonymous SNP	64699203 bp, A/G
Sobic.006G253900	6	*SbF3H*	Non-synonymous SNP	59157048 bp, A/T; 59157274 bp, C/T; 59158255 bp, T/A
Sobic.006G254000	6	*SbF3H*	Non-synonymous SNP	59160879 bp, A/C; 59161461 bp, G/A
Sobic.007G047300	7	*SbOMT*	Non-synonymous SNP	4721737 bp, G/C; 4721966 bp, C/T; 4724116 bp, T/C
Sobic.010G052200	10	*SbOMT*	Non-synonymous SNP	4072017 bp, C/G
Sobic.008G030100	8	*SbCHI*	Non-synonymous SNP	2684008 bp, C/G

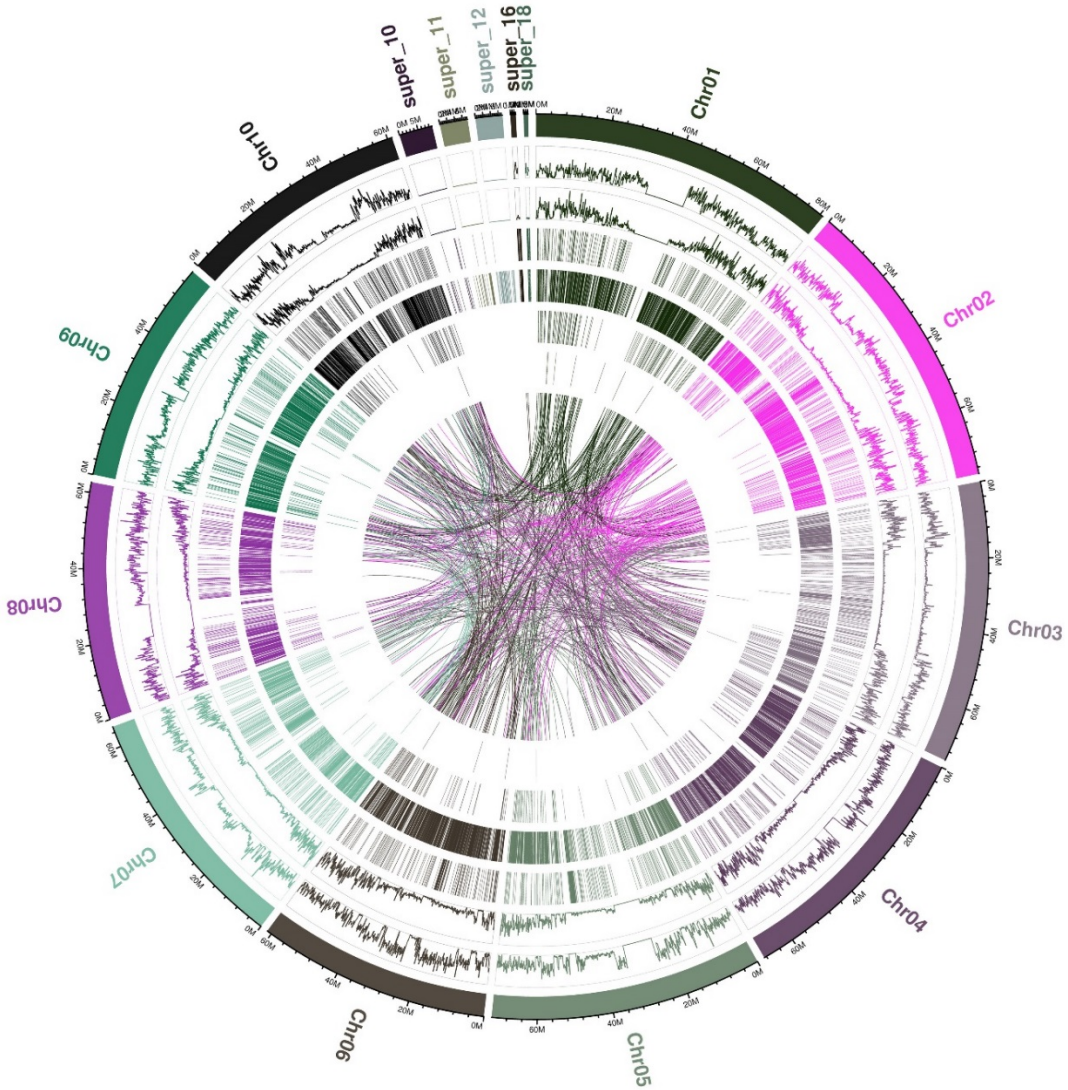


Fig. 1. Genome-wide landscape of genetic variation in Hongyingzi. Cycles from outside to inside indicate chromosome, SNP, Indel, CNV duplication, CNV deletion, SV insertion, SV deletion, SV invertion, SV ITX, and SV CTX. **ITX**: Intrachromosomal translocation. **CTX**: Interchromosomal translocation.

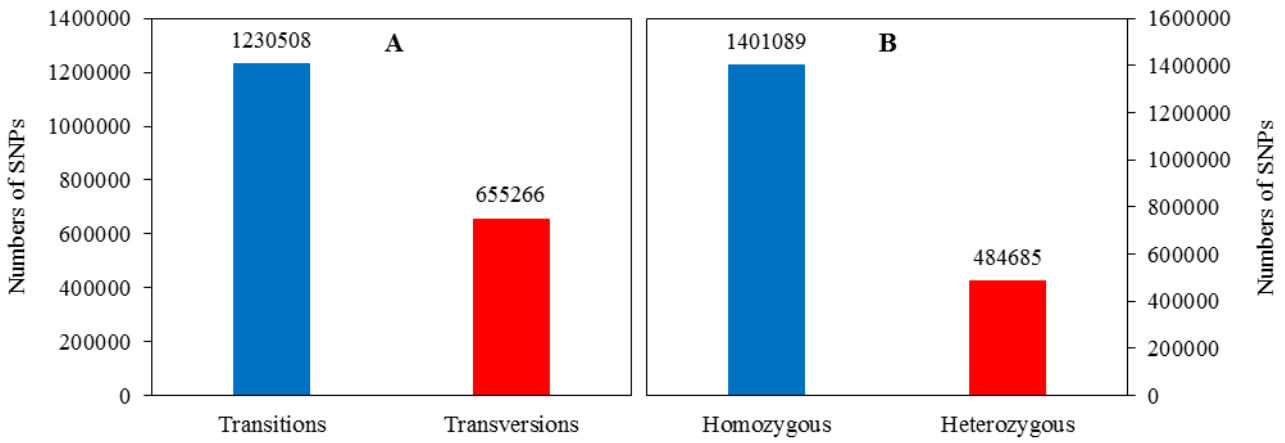


Fig. 2. SNP distribution in the Hongyingzi genome. **A**: Numbers of transitions and transversions SNPs. **B**: Numbers of homozygous and heterozygous SNPs.

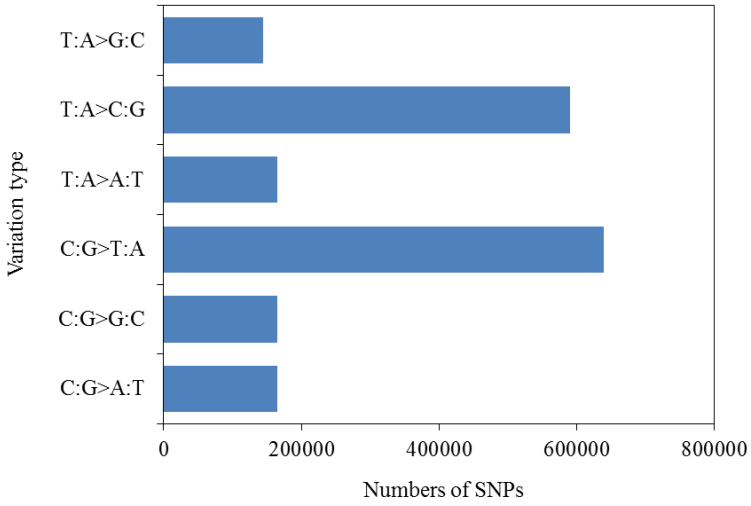


Fig. 3. Distribution of SNP variation types

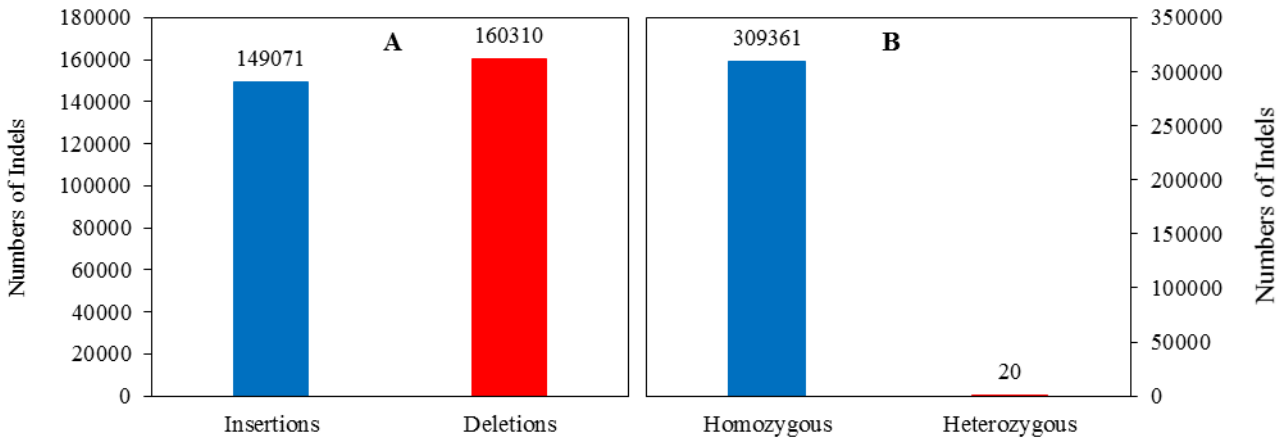


Fig. 4. Indel distribution in the Hongyingzi genome. **A**: Numbers of insertions and deletions. **B**: Numbers of homozygous and heterozygous Indels.

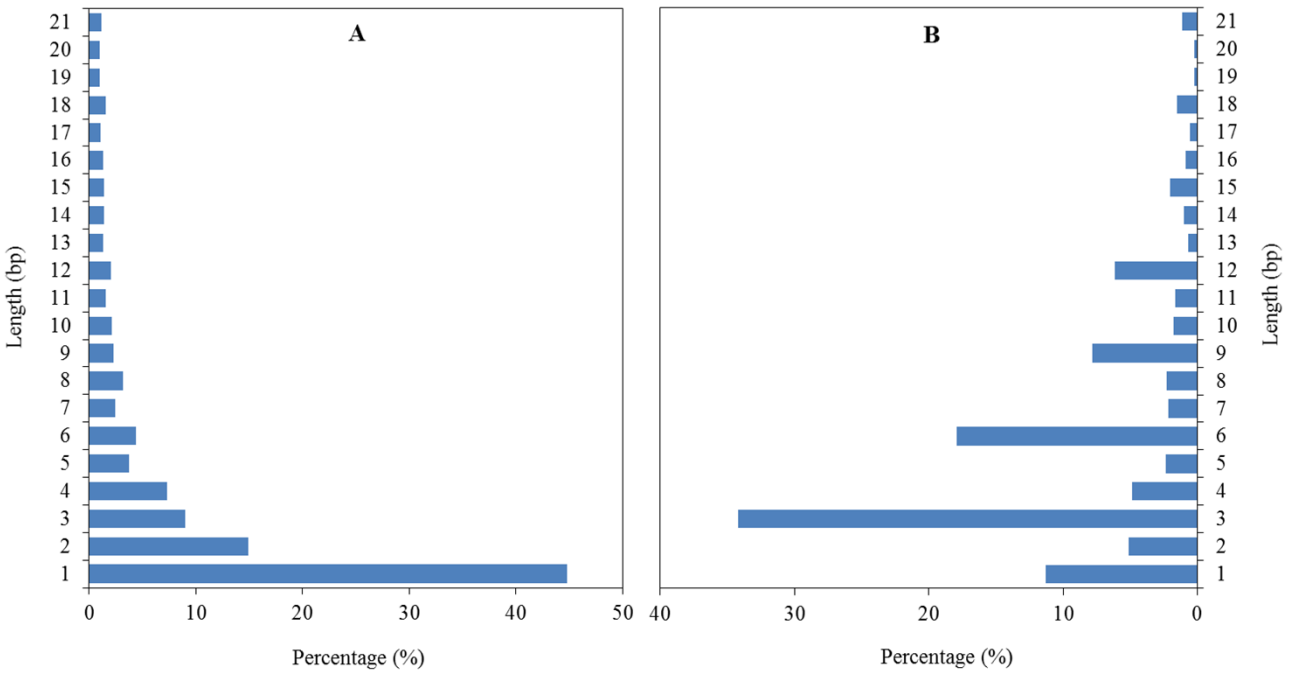


Fig. 5. Length distribution of Indels in whole genome and coding regions. **A**: Length distribution of Indels in whole genome. **B**: Length distribution of Indels in coding regions.

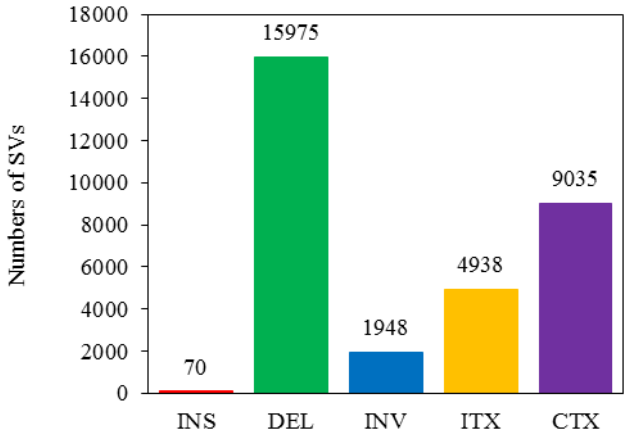


Fig. 6. SV distribution in the Hongyingzi genome. **INS**: Insertions. **Del**: Deletions. **INV**: Inversions. **ITX**: Intrachromosomal translocations. **CTX**: Interchromosomal translocations.

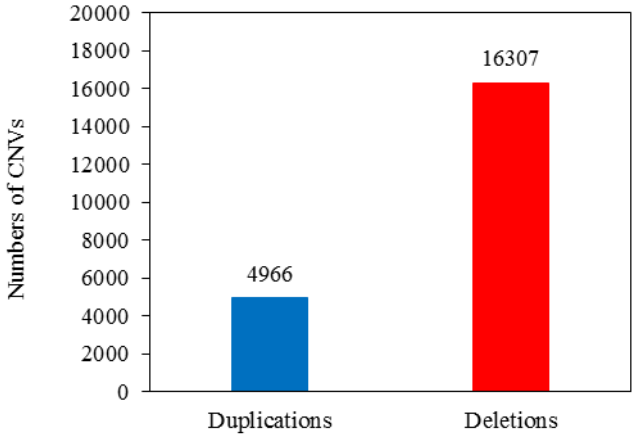


Fig. 7. CNV distribution in the Hongyingzi genome.

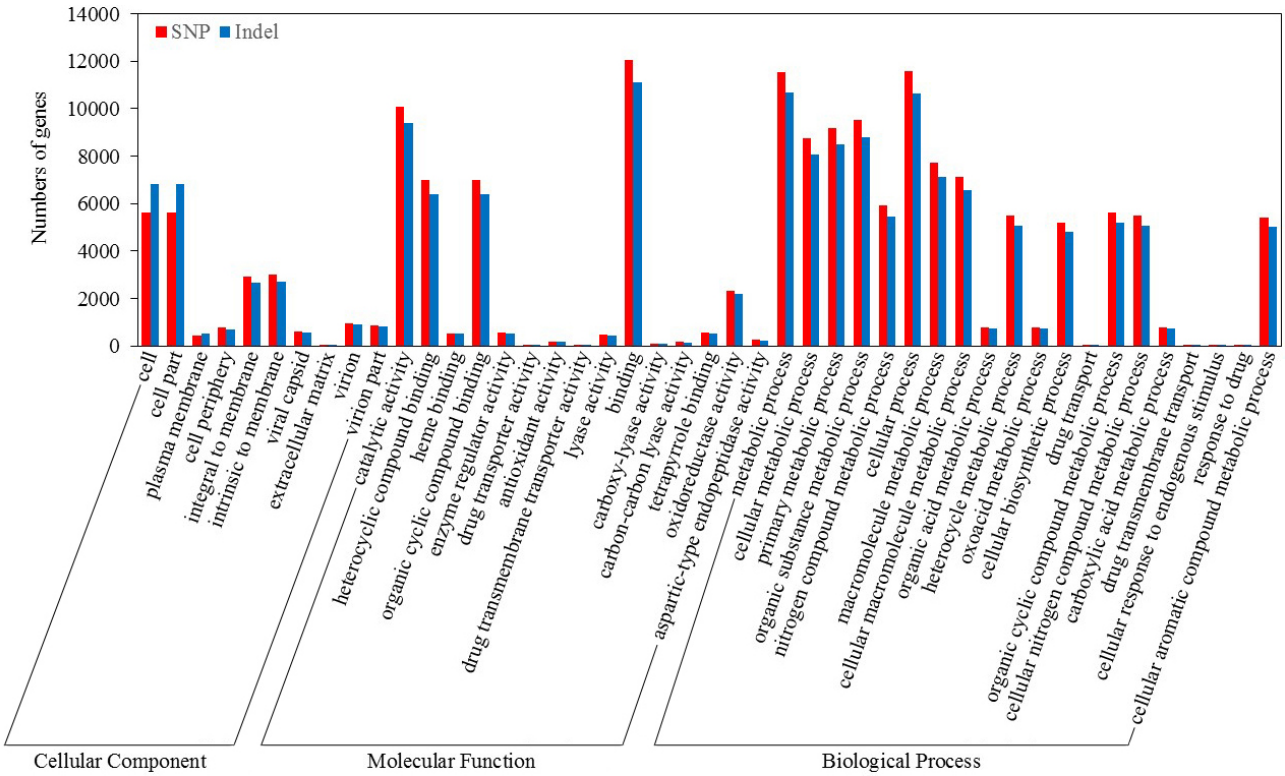
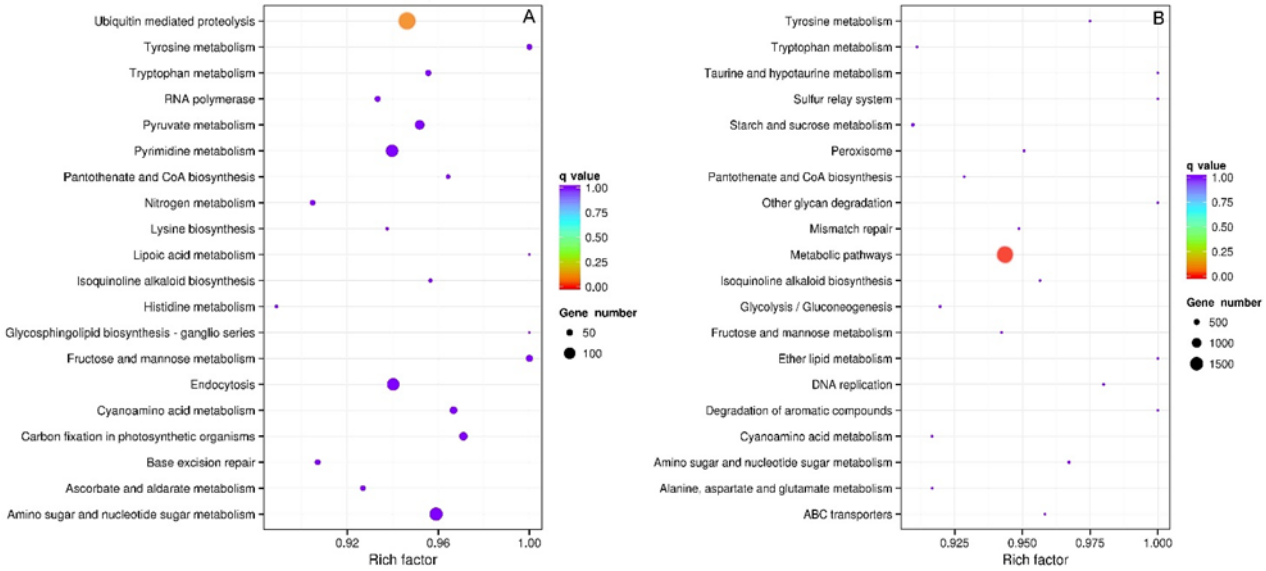


Fig. 8. Classification of gene variations compared with GO database.

905



906

907 **Fig. 9.** Classification of gene variations compared with KEGG database. **A**: Gene variations caused by SNPs. **B**: Gene variations caused

908 by Indels.