Indian Journal of Agricultural Sciences **89** (7): 1121–5, July 2019/Article

De-novo transcriptome assembly for discovery of putative microsatellite markers and transcription factors in black pepper (*Piper nigrum*)

ANKITA NEGI[1], RAHUL SINGH JASROTIA[2], SARIKA JAISWAL[3], U B ANGADI[4], M A IQUEBAL[5], JOHNSON GEORGE K[6], ANIL RAI[7] and DINESH KUMAR[8]

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India
ICAR-Indian Institute of Spices Research, Kozhikode, Kerala 673 012, India

Received: 15 January 2019; Accepted: 22 January 2019

ABSTRACT

Black pepper (*Piper nigrum* L.) (2n= 52; *Piperaceae*), is a perennial, trailing woody flowering vine having global importance with widespread dietary, medicinal, and preservative uses. It is one of the highly traded high cost spice germplasms cultivated for its fruit. Unlike model species, the whole genome sequence information and genomic resources of black pepper are still unavailable in public domain. Crop production is highly affected by abiotic and biotic stresses. Hence transcriptome profiling has permitted a significant enhancement in discovery and expression profiling of genes and functional genomic studies in black pepper. Stress responsive transcriptomic data of various black pepper genotypes were obtained from public domain (SRA database, NCBI) for the *de novo* transcriptome assembly, identification of transcription factors and mining of putative simple sequence repeat markers. *De novo* transcriptome assembly was done with SOAP *denovo-trans* assembler, which generated 53690 transcripts. A total of 14005 transcription factors with BLAST and 39685 without BLAST hits were identified. A total of 4770 putative SSR markers were identified using de novo transcriptome assembly. Myeloblastosis (*MYB*) related proteins, Basic helix-loop-helix (*BHLH*), NAC, WRKY and ERF transcriptional factors found in this study are reported to be associated with plant tolerance against stress condition. These SSR markers can be valuable and facilitate advancements in genetic and molecular studies in the endeavour of better productivity of *P. nigrum* germplasm, especially in the era of rising abiotic stress.

Key words: Black pepper, *de novo* assembly, Markers, RNA-sequence, Transcription factors

Black pepper (*Piper nigrum* L.) (2n = 52), a trailing woody flowering vine belongs to *Piperaceae* family. Genus *piper* has more than 1000 species (Jaramillo and Manos 2001). It is categorized into different types based on the degree of maturation and type of processing method used. It is mainly a self-pollinated plant and cultivated commercially by orthotropic stem cutting method (Krishnamoorthy and Va 2011). It is known as "*King of spices*" due to its global trade, widespread dietary, medicinal, preservative and insecticidal uses (Quijano-Abril *et al.* 2006). It has great nutritional and agricultural significance with antioxidant, anti-inflammatory and anticancerous properties (Gulcin I. 2005).

Black pepper is known to have originated in tropical evergreen forests of Western Ghats of India. It is one of the highly traded spices of the world and cultivated as a major cash crop in more than 30 tropical countries of the world (Ahmad *et al.* 2010, Tian *et al.* 2006). Vietnam is the world's leading and largest producer and exporter of pepper, producing about 35% of the world's *P. nigrum* crop. Globally, India contributes second highest area of black pepper (132000 ha, 2017) after Indonesia (181978 ha, 2017) (FAO 2017).

The productivity of the crop is affected by both biotic as well as abiotic stresses. Since, the crop is one of the costliest spice germplasms, these stresses lead to major economic losses. Currently, a few studies on this crop transcriptome/ RNA-Seq has been done. Also, the whole genome sequencing of black pepper is yet unavailable.

The continuous global rise in temperature and loss in the crop production due to abiotic stresses warrants the study on pathways and its mechanisms involved in abiotic stress tolerance. This will be useful for further investigation to elucidate stress improvement, deciphering pathways as well as mining of genic region SSR markers along with the primer generation which will further be useful in QTL mapping population and breeding programmes for germplasm improvement.

Present address: [1]Research Scholar (negiwinx@gmail.com), [2]Research Scholar (rahuljasrotia86@gmail.com), [3]Senior Scientist (sarika@icar.gov.in), [4]Principal Scientist (angadiub@iasri.res.in), [5]Senior Scientist (ma.iquebal@icar.gov.in), [7,8]Principal Scientist (Anil.Rai@icar.gov.in, dinesh.kumar@icar.gov.in), ICAR-IARI New Delhi; [6]Principal Scientist (kokkatjohn@rediffmail.com), ICAR-IISR, Kerala.

MATERIALS AND METHODS

Data Set: Single end and paired end transcriptomic SRA data of *P. nigrum* were obtained from different genotypes available at National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov). *SRAtoolkit* was used to convert SRA data into Fastq format, resulting in separate files for forward and reverse data for each sample.

Pre-processing and de novo assembly: FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Andrews 2010) was used for the visualization of reads quality, before and after the pre-processing. Quality of the data was based on various parameters such as basic statistics, per sequence quality scores, per base sequence content, adapter content, per sequence GC content, sequence length distribution, per base N content, sequence duplication levels, over-represented sequences, per tile sequence quality and *K*-mer content. Trimmomatic tool (version 0.36) was used for removal of low quality reads (Bolger *et al.* 2014). Trimming of bases were done from 3' and 5' end, keeping headcrop as 10-12 and phred-score of 33. These high quality pre-processed reads were used for *de novo* transcriptome assembly using SOAP*denovo-Trans* (version 0.99) assembler (Xie *et al.* 2014) that provides higher contiguity, faster execution and lower redundancy, which not only removes sequencing errors but also shortens ambiguous contigs (default ≤100 bp) caused by repeats. This was followed by CAP3 assembler for removal of redundant sequences (Huang and Madan 1999).

Homology Search, Annotation and Functional Characterization: Homology search of transcripts from *P. nigrum* transcriptome assembly were performed against NCBI non-redundant database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) using Blastx algorithm as standalone local ncbi-blast-2.2.31+ with threshold E-value 1e-3 (Altschul *et al.,* 1990). For further research, Blast2GO tool (https://www.blast2go.com/) (Conesa *et al.,* 2005) was used for their mapping and annotation. The functional classification was done using the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases for broader overview of the crop species and the pathways involved.

Prediction of transcription factors involved in stress tolerance: Transcriptional factors from the generated assembly were predicted using PlantTFDB 4.0 (http://planttfdb.cbi.pku.edu.cn/download.php) (Jin *et al.* 2016). Blastx algorithm as standalone local ncbi-blast-2.2.31+ with threshold E-value 1e-6 (Altschul *et al.* 1990) were employed.

Mining of genic region putative molecular markers and primer designing: Mining of genic region putative SSR markers was done from *de novo* transcriptome assembly of *P. nigrum* using perl script of MISA (Micro SAtellite identification tool) (http://pgrc.ipk-gatersleben.de/misa/) (Thiel *et al.* 2003). The SSR loci containing repeat units of 1–6 nucleotides only were considered. For identification of genic region SSR markers, default parameters like ten repeating units for mononucleotides, six repeating units for dinucleotides and five repeating units for trinucleotides, tetranucleotides, pentanucleotides and hexanucleotides were considered. Maximum difference between two SSRs was kept as 100 bp. SSR specific primers were designed using Primer3 V0.4.0 (Untergasser *et al.* 2012).

RESULTS AND DISCUSSION

Pre-processing and de novo assembly: Entire public domain available genomic resource of species *P. nigrum* were mined successfully to discover transcription factors, molecular markers and to perform functional annotation. We retrieved 5 sets of single end and 4 sets of paired end transcriptomic SRA data of *P. nigrum* L. from NCBI (Table 1). A total of 516057410 single end and 157856524 paired end reads of black pepper genotype samples were retrieved with reads length 101 bp. After data pre-processing using Trimmomatics and visualization by FastQC, a total of 52545628 and 967085 low quality reads from single end and paired end samples, respectively, were removed. The remaining high quality reads were then used for the downstream analysis. The *de novo* transcriptome assembly by SOAP *de novo-trans* generated a total of 53690 transcripts with percent GC as 43.82% and GC count 14993012 bp. The N50 value of the contig was 688 bp.

Homology Search: Homology search of transcripts were performed using blastx to identify the similar genes present in the database. Out of 53690 transcripts, we found that 34996 transcripts showed similarity with other genes present in the database, while 15894 transcripts were novel as they didn't showed any similarity or were without hits. A total of 639 transcripts were involved in mapping. Top hit species distribution revealed that maximum hits were found with *Nelumbo nucifera* i.e. 4938 transcripts, followed by *Macleaya cordata* and *Elaeis guineensis* in 3098 and 1597 transcripts, respectively. Total transcripts were found to be involved in 111 pathways. Maximum number of transcripts were found to be involved in biosynthesis of antibiotics i.e. 46, followed by 15 and 13 transcripts in purine metabolism and Glycolysis / Gluconeogenesis, respectively (Fig1). Blast2GO annotation results revealed that these transcripts gets categorized into three sub categories such as biological process, molecular functions and cellular components. In biological process, molecular function and cellular

Table 1. List of transcriptome data retrieved for analysis (PE: Paired End; SE: Single End)

S.No.	Accession IDs	Description
1	SRR1776865	Vegetative phase plantlets (PE)
2	SRR1777719	Root transcriptome(PE)
3	SRR1781514	Root transcriptome(PE)
4	SRR408047	Leaf transcriptome(PE)
5.	SRR1164727	SE
6.	SRR1818148	Fruit transcriptome (SE)
7.	SRR1583631	Root transcriptome (SE)
8.	SRR3341858	Root transcriptome (SE)
9.	SRR3341859	Root transcriptome (SE)

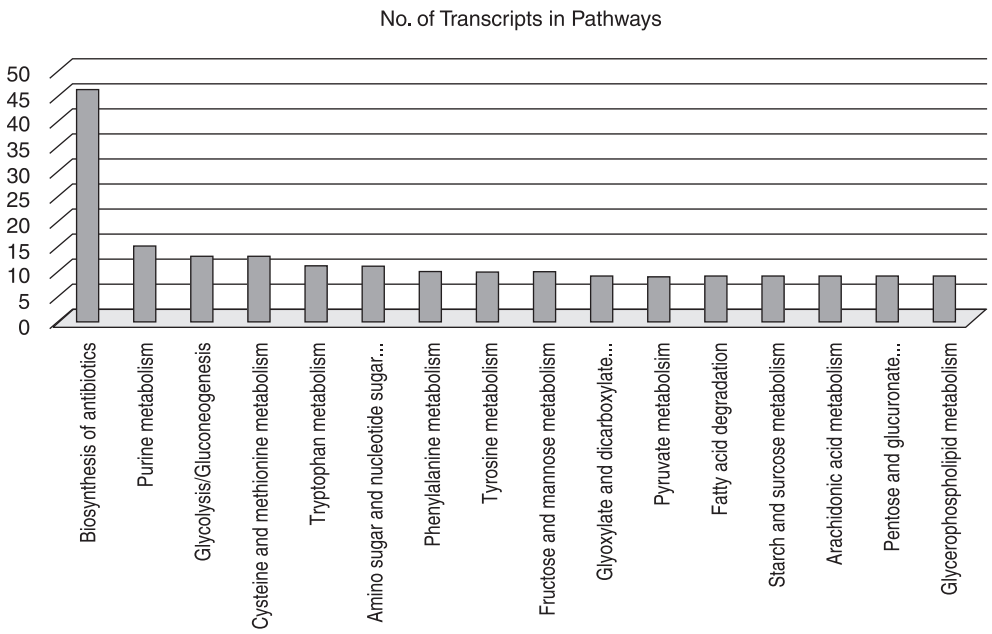

Fig 1. Top 16 KEGG pathways of transcripts of *P. nigrum.*

component, a total of 10230, 2316 and 6096 transcripts were found, respectively (Fig 2).

Prediction of transcription factors involved in stress tolerance: Since transcription factors are major players in development and adoptive responses in abiotic stress, thus they also play major role in putative candidate genes involved in transcriptional regulation of abiotic stress mediation. Transcription factors (TF) were identified by performing blast (Blastx tool) against the plant transcription factor database (PlantTFDB 4.0) using blastx tool. A total of 14005 transcription factors among 53690 transcripts and 39685 transcripts were novel as they were without hits with expected e-value 1e-6. Most abundant transcription factors were bHLH, MYB, NAC, ERF, C2H2 and WRKY represented by 1470, 1418, 1089,783, 672 and 620 transcripts, respectively. Top fifteen most

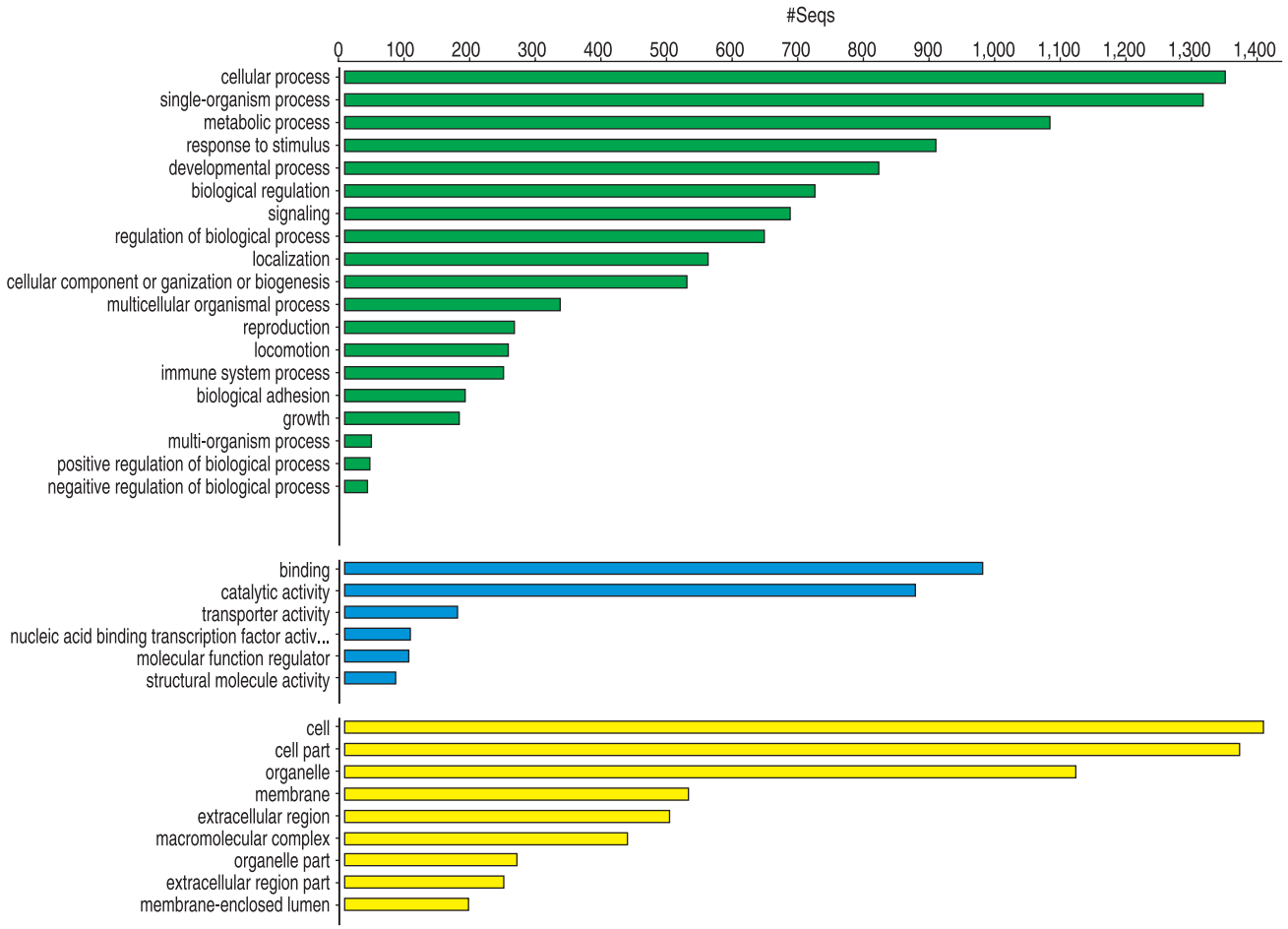

Fig 2. Gene ontology of transcripts. (Green, blue and yellow colour represent the biological process, molecular functions and cellular components, respectively).

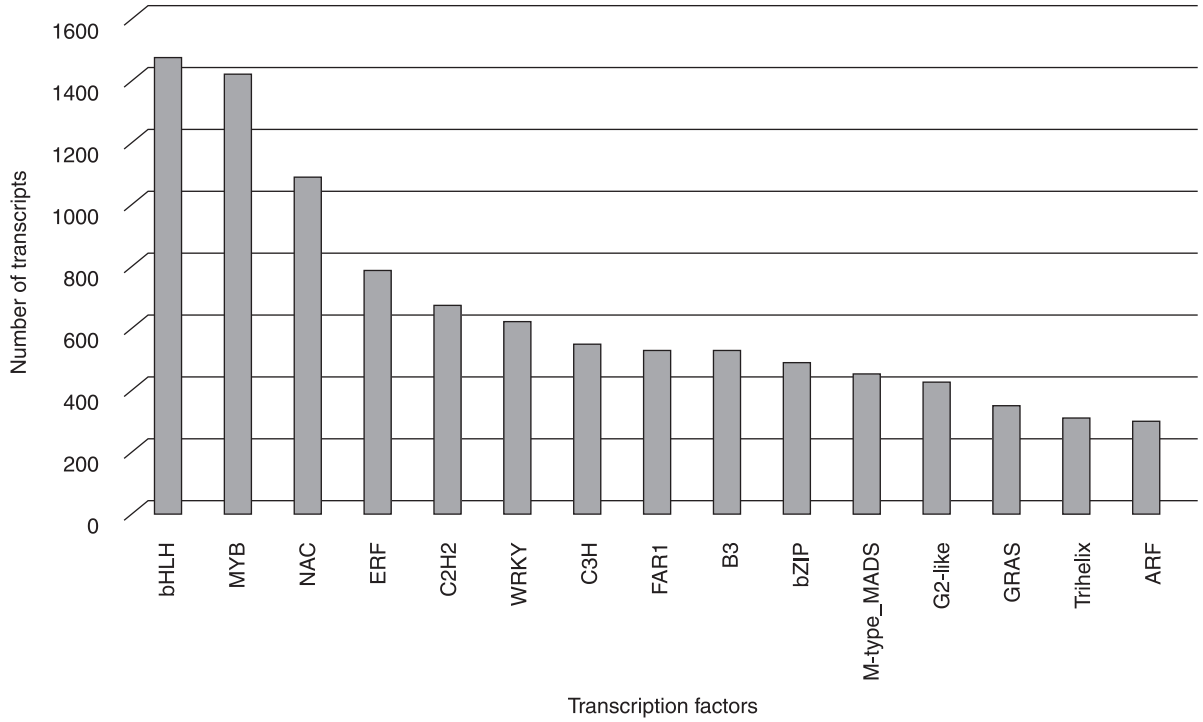

Fig 3. Top 15 transcription factors identified from transcripts.

abundant transcription factors identified from transcripts are represented in Fig 3.

Among the discovered TFs, basic helix-loop-helix (bHLHs) is well known regulator of abiotic defence mechanism. It activates the different types of genes which are involved in sensing of environmental signals by plant like hormone signalling. Myeloblastosis related proteins (MYB) is a huge and diverse family and is found mostly in all eukaryotes and are known to be highly expressed in drought. They are found to be involved in different processes, such as the control of cellular and organ morphogenesis, circadian rhythm, secondary metabolism as well as regulation of stomatal movements as a response to drought stress (Shin *et al.* 2011). C2H2 zinc-finger (C2H2-ZF) proteins are a large gene family in plants that participate in biotic and abiotic stress responses as well as various aspects of normal plant growth and development. This domain has also shown to mediate RNA, protein interactions and is known to be involved in overlapping responses to a variety of stress conditions including environmental stress regulation (Liu *et al.* 2015). TFs from other families such as HD-Zip (homeodomain-leucine zipper), GRAS (GAI – Gibberellin-acid insensitive), RGA – Repressor of GA1, SCR - Scarecrow), HSF (heat shock factor) and NF-Y (nuclear factor Y) which are found in our study, have been known to cope not only with abiotic stresses like salt, temperature and drought stresses but also with other stresses such as heat shock, oxygen deficiency, high light and nutrient deficiency (Lan *et al.* 2017). Members of WRKY domain protein family have also been found in our study which are known to contain at least one conserved DNA-binding region including highly conserved a zinc finger motif ($CX_{4-7}CX_{22-23}HXH/C$) and WRKYGQK peptide sequence (Pandey and Somssich 2009).

Mining of genic region putative molecular markers and primer designing: A total of 4770 SSRs were mined from the *de novo* transcriptome assembly of *P. nigrum,* Out of these 355 transcripts were with more than one SSR loci. There was less abundance (135) of compound SSR. Tri-nucleotide repeats were most abundant (1407) followed by mono-type (2573) and di nucleotide (754). Being coding region, these transcripts are expected to have higher abundance of tri-nucleotide repeats (Huang *et al.* 2016) (Table 2). In order to use the discovered SSR loci, primers were computed using PRIMER3 tool and 382 primer pairs were obtained successfully. These are ready to use primers for genotyping which requires wet-lab validation.

Mining of SSR markers from transcriptomic data can cater the need of crisis of molecular markers having advantage in terms of time and cost effectiveness. SSR serves as the most versatile molecular markers. In the present study, we discovered SSR loci from genic regions using transcriptomic data which offers several advantages in terms of stability and transferability. These markers can be used in linkage mapping, studies related to genetic variability and functional diversity (Kujur *et al.* 2013). Applications of such markers are evident in crops like tomato and pepper, sugarcane, basil, sesame, African oil palm and tea (Taxak *et al.* 2017). These SSR primers could represent a valuable and useful genomic resource of *P. nigrum which w*ill facilitate further advancements in genetic and molecular studies in the endeavour of better productivity of *P. nigrum* germplasm, especially in the era of rising abiotic stress.

Table 2 Summary statistics of mined genic region putative SSRs and primers from *P. nigrum*

Information regarding SSRs	*De- novo* Assembly
Sequences examined	53690
Identified SSRs	4770
SSR containing sequences	4377
Sequences containing >1 SSR	355
SSRs present in compound Formation	135
Mono- nucleotide	2573
Di- nucleotide	754
Tri- nucleotide	1407
Tetra- nucleotide	32
Penta- nucleotide	3
Hexa- nucleotide	1
Total no. of primers	382

ACKNOWLEDGEMENTS

The authors are thankful to Indian Council of Agricultural Research (ICAR), Ministry of Agriculture and Farmers' Welfare, Government of India for providing financial and infrastructural support to carry out this research and for creation of Advanced Super Computing Hub for Omics Knowledge in Agriculture (ASHOKA) facility where the work was carried out. The CAB in grant (Grant Number: F. no. Agril. Edn.4-1/2013-A&P) by ICAR is also acknowledged. The authors further acknowledge the supportive role of Directors of ICAR-IASRI, New Delhi and ICAR-IISR, Kerala. The grant of Junior Research Fellowship to AN by Indian Council of Agricultural Research is duly acknowledged.

REFERENCES

Ahmad N, Fazal H, Abbasi B H, Rashid M, Mahmood T and Fatima N. 2010. Efficient regeneration and antioxidant potential in regenerated tissues of *Piper nigrum* L. *Plant Cell, Tissue and Organ Culture (PCTOC)* **102**(1): 129–34.

Altschul S F, Gish W, Miller W, Myers E W and Lipman D J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**(3): 403–10.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.

Bolger A M, Lohse M and Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15): 2114–20.

Conesa A, Gotz S, Garcia-Gomez J M, Terol J, Talon M and Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18): 3674–6.

FAOSTAT: Production, Crops, Millet, data. Food and Agriculture Organization. 2017.

Gulcin I. 2005. The antioxidant and radical scavenging activities of black pepper (*Piper nigrum*) seeds. *International Journal of Food Sciences and Nutrition* **56**(7): 491–9.

Huang X and Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Research* **9**(9): 868–77.

Huang X, Yan H D, Zhang X Q, Zhang J, Frazier T P, Huang D J, Lu L,Huang L, Liu W, Pneg Y, Ma X D and Yan Y. 2016. *De novo* transcriptome analysis and molecular marker development of two hemarthria species. *Front Plant Science* **7**: 496.

Jaramillo M A and Manos P S. 2001. Phylogeny and patterns of floral diversity in the genus Piper (Piperaceae). *American Journal of Botany*, **88**(4): 706–16.

Jin J, Tian F, Yang D C, Meng Y Q, Kong L, Luo J and Gao G. 2016. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research* **45**(D1): 1040–45

Krishnamoorthy B and Parthasarathy V A. 2011. Improvement of black pepper. *Plant Sciences Reviews* 2010, 37.

Kujur A, Bajaj D, Saxena M S, Tripathi S, Upadhyaya H D, Gowda C L L, Singh S, Jain M, Tyagi A K and Parida S K. 2013. Functionally relevant microsatellite markers from chickpea transcription factor genes for efficient genotyping applications and trait association mapping. *DNA Research* **20**(4): 355–74.

Lan Thi Hoang X, Du Nhi N H, Binh Anh Thu N, Phuong Thao N and Phan Tran L S. 2017. Transcription factors and their roles in signal transduction in plants under abiotic stresses. *Current genomics* **18**(6): 483–497.

Liu Q, Wang Z, Xu X, Zhang H and Li C. 2015. Genome-wide analysis of C2H2 Zinc-finger family transcription factors and their responses to abiotic stresses in poplar (*Populus trichocarpa*). *PloS One* **10**(8): e0134753.

Pandey S P and Somssich I E. 2009. The role of WRKY transcription factors in plant immunity. *Plant Physiology* **150**(4): 1648–55.

Quijano-Abril M A, Callejas-Posada R and Miranda-Esquivel D R. 2006. Areas of endemism and distribution patterns for neotropical *Piper* species (Piperaceae). *Journal of Biogeography* **33**(7): 1266–78.

Shin D, Moon S J, Han S, Kim B G, Park S R, Lee S K and Yi B Y. 2011. Expression of StMYB1R-1, a novel potato single MYB-like domain transcription factor, increases drought tolerance. *Plant Physiology* **155**(1): 421–32.

Taxak P C, Khanna S M, Bharadwaj C, Gaikwad K, Kaur S, Chopra M, Tandon G, Jaiswal S, Iquebal M A, Rai A, Kumar D, Srinivasan and Jain P K. 2017. Transcriptomic signature of fusarium toxin in chickpea unveiling wilt pathogenicity pathways and marker discovery. *Physiological and Molecular Plant Pathology* **100**: 163–77.

Thiel T, Michalek W, Varshney R and Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare L.*). *Theoretical and Applied Genetics* **106**(3): 411–22.

Tian B, Lin Z B, Ding Y and Ma Q H. 2006. Cloning and characterization of a cDNA encoding Ran binding protein from wheat: Full Length Research Paper. *DNA Sequence* **17**(2): 136–42.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth B C, Remm M and Rozen S G. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Research* **40**(15): 115.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S and Zhou X. 2014. SOAP *denovo-Trans*: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**(12): 1660–6.