



**Full Length Article**

# Identification, Phylogenetic Evolution and Expression Analysis of GATA Transcription Factor Family in Maize (*Zea mays*)

Long Jiang, Xiaoming Yu\*, Dianyuan Chen\*, Hu Feng and Jianming Li

College of Agronomy, Jilin Agricultural Science and Technology University, Jilin, Jilin 132101, Jilin Province, P.R. China

\*For Correspondence: [jljnykjxyjl@163.com](mailto:jljnykjxyjl@163.com); [yuxm446@nenu.edu.cn](mailto:yuxm446@nenu.edu.cn)

Received 30 September 2019; Accepted 26 November 2019; Published 04 February 2020

## Abstract

GATA transcription factors widely exist in eukaryotes and play important role in the regulation of plant biological processes. The present study identified the transcription factors of GATA in maize by bioinformatics method, and 38 GATA genes and 30 GATA genes in *Arabidopsis*. Phylogenetic analysis of GATA genes in maize and *Arabidopsis* showed that family II was unique to maize. In conservative domain analysis, 38 GATA transcription factors in maize all contain GATA structure, among which GATA-12, GATA-13, GATA-16, GATA-31 and GATA-34 contain other structures. The expression patterns of GATA transcription factor family genes were different in different maize tissues and developmental stages, and 10 transcription factors were not expressed during the analysis period; 9 transcription factors expressed at high levels in all stages and tissues during the analysis period; and 8 transcription factors were present medium level expression in all stages and tissues. GATA-17 was expressed at a high level in coleoptile, germination and internode, GATA-6 and GATA-7 were expressed at a specific level in leaves. Subcellular localization predicted that all 38 GATA transcription factor families were in the nucleus, indicating their role in regulating transcription. Thus, identification, phylogenetic tree, chromosome localization and expression pattern of GATA transcription factors in maize will provide basis for further studies. © 2020 Friends Science Publishers

**Key words:** Maize; GATA transcription factor; Physical and chemical properties; Chromosome location; Expression analysis

## Introduction

Maize (*Zea mays* L.) is the most important food, feed, industrial raw materials and energy crops in the world. It plays an important role in ensuring world food security, economic development and alleviating the energy crisis (Zhao *et al.* 2016, 2018). With the continuous improvement of people's living standards in China, the demand for maize is increasing sharply. Therefore, increasing maize yield has a long-term significance to promote the safe and healthy development of grain in China.

GATA transcription factors are a kind which widely exist in eukaryotes (Shi and Xu 2005) and belong to zinc finger transcription factors. They play an important role in the regulation of plant photoresponse, chlorophyll synthesis, cytokinin response, carbon and nitrogen metabolism, and in the regulation of plant flowering time, leaf extension growth, flower development and other biological processes (Yuan *et al.* 2017a). At present, reports on GATA transcription factors in maize are rare, so it is necessary to analyze the transcription factor family and further understand the role of GATA transcription factors. In present study, the amino acid sequence, chromosome localization, physico-chemical properties and subcellular structure of

GATA transcription factors in maize were predicted and analyzed by bioinformatics. Bioinformatics analysis of the GATA family in maize can lay a foundation for further understanding the function of GATA transcription factors in maize, and provide a theoretical basis for further maize breeding and maize yield using GATA transcription factors.

## Materials and Methods

### Test material

The sequence of nucleic acid and protein of maize and *Arabidopsis thaliana*, GATA transcription factor family gene were downloaded from PlantTFDB (<http://planttfdb.cbi.pku.edu.cn/>) (Liu *et al.* 2015).

### Acquisition of GATA transcription factor sequence

Using the database Plant TFDB (<http://planttfdb.cbi.pku.edu.cn/>), the amino acid sequences of GATA transcription factor family proteins in *Arabidopsis thaliana* and maize were downloaded, 30 and 38 amino acid sequences of GATA transcription factor family proteins were obtained, respectively. The sequence of GATA proteins

in maize and *Arabidopsis thaliana* obtained from the above databases includes proteins with multiple transcription versions of GATA gene translation. In this study, the longest selection transcripts (and their encoded proteins) of GATA gene with multiple transcripts were subsequently analyzed.

### Construction of phylogenetic tree of transcription factor GATA gene

ClustalX was used to sequence 38 maize GATA transcription factor family proteins and 30 *Arabidopsis thaliana* GATA transcription factor family proteins. Neighbor Joining (NJ) method built in MEGA5.05 was used to construct the adjacency tree. Bootstrap was repeated 1000 times, and other parameters were set by default (Zhao *et al.* 2015).

### Analysis of conserved domains of amino acid sequences of GATA transcription factor family proteins

Using Pfam Domain pattern redrawer function in TBtools software, the conserved domains of amino acid sequences of GATA proteins in maize were predicted.

### Location of GATA gene on chromosome

According to the published information of maize genome B73 RefGen\_v3, the distribution of 38 GATA transcription factor family genes on 10 maize chromosomes was determined. The location of GATA gene on chromosome was found by BLSAT comparison in Maize GDB database (<https://www.maizegdb.org/>).

### Analysis of physicochemical properties of amino acids

On-line Protparam software (<http://web.expasy.org/protparam/>) provided by ExPaSy was used to analyze the number, molecular weight, theoretical isoelectric point, aliphatic amino acid number and protein hydrophobicity (Xie *et al.* 2014).

### Sub-cellular localization of GATA

Plant-mPLoc (<http://www.csbio.sjtu.edu.cn/bioinf/Plant-multi/>), the location of 38 GATA transcription factor family proteins in cells was predicted (Zhu *et al.* 2012).

### Expression analysis of GATA family genes

The expression patterns of 38 GATA transcription factor family genes in different maize tissues and developmental stages were analyzed using published transcriptome data. Drawn thermal maps through the amazing heatmap function in ttools tools.

## Results

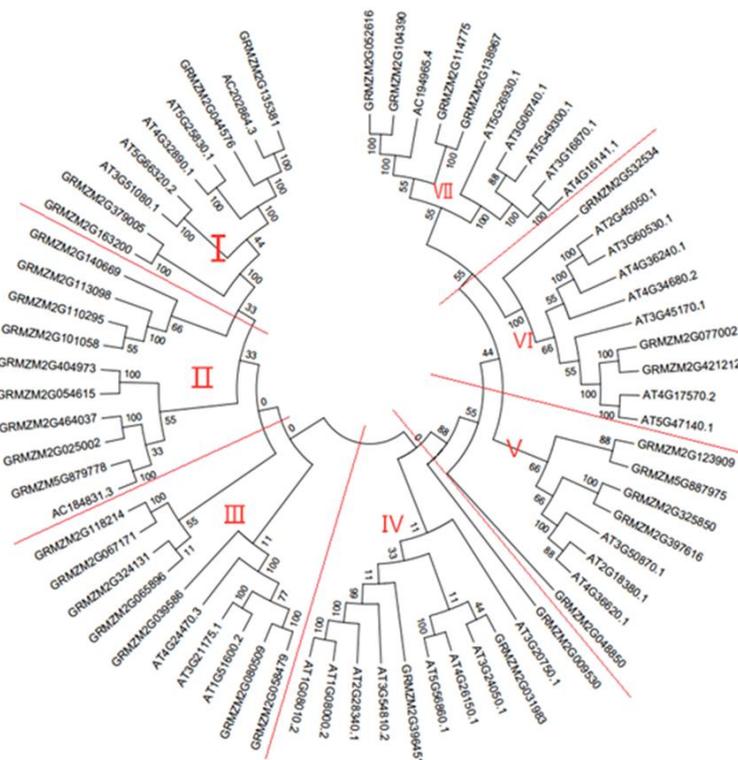
### Identification of GATA transcription factors

According to the plant transcription factor database

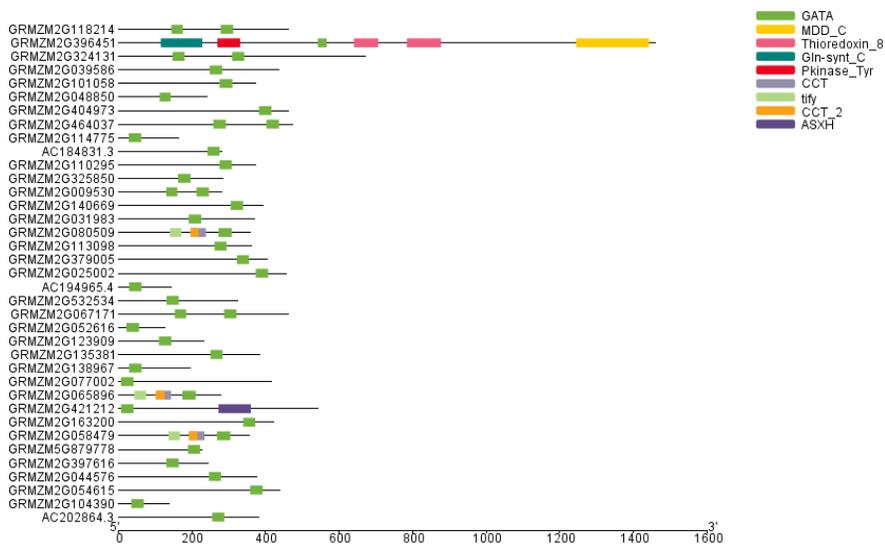
PlantTFDB (<http://planttfdb.cbi.pku.edu.cn/>), the GATA proteins identified were analyzed. For the proteins translated from different transcripts with the same GATA gene, the protein with the longest amino acid sequence was selected for the following analysis. After screening, 30 GATA proteins from *Arabidopsis thaliana* and 38 GATA proteins from maize were obtained: AC184831, AC194965, AC202864, GRMZM2G009530, GRMZM2G025002, GRMZM2G031983, GRMZM2G039586, GRMZM2G044576, GRMZM2G048850, GRMZM2G052616, GRMZM2G054615, GRMZM058479, GRMZM2G065896, GRMZM2G067171, GRMZM2G077002, GRMZM2G080509, GRMZM2G101058, GRMZM2G104390, GRMZM2G110295, GRMZM2G113098, GRMZM2G114775, GRMZM2G118214, GRMZM2G123909, GRMZM2G135381, GRMZM2G138967, GRMZM2G140669, GRMZM2G163200, GRMZM2G324131, GRMZM2G325850, GRMZM2G379005, GRMZM2G396451, GRMZM2G397616, GRMZM2G404973, GRMZM2G421212, GRMZM2G464037, GRMZM2G532534, GRMZM5G879778, GRMZM5G887975. Thirty-eight GATA genes in maize were named respectively GATA-1, GATA-2, GATA-3, GATA-4, GATA-5, GATA-6, GATA-7, GATA-8, GATA-9, GATA-10, GATA-11, GATA-12, GATA-13, GATA-14, GATA-15, GATA-16, GATA-17, GATA-18, GATA-19, GATA-20, GATA-21, GATA-22, GATA-23, GATA-24, GATA-25, GATA-26, GATA-27, GATA-28, GATA-29, GATA-30, GATA-31, GATA-32, GATA-33, GATA-34, GATA-35, GATA-36, GATA-37, GATA-38.

### Phylogenetic Analysis of GATA Gene in Maize and *Arabidopsis thaliana*

Thirty GATA genes from *Arabidopsis thaliana* and 38 GATA genes from maize were sequenced by ClustalW. The GATA transcription factor family of maize and *Arabidopsis* can be divided into seven subfamilies, which are named I, II, III, IV, V, VI (Fig. 1). It was found that subfamily II is unique to maize. Other subfamily I, III, IV, V, VI, \_maize and *Arabidopsis* proteins exist. The number of GATA proteins in subfamily I, III, IV, V, VI maize and *Arabidopsis* are different. For example, subfamily I contains 5 maize GATA and 4 *Arabidopsis* GATA. Subfamily III contains seven maize GATA and three *Arabidopsis* GATA, subfamily IV contains three maize GATA and eight *Arabidopsis* GATA, subfamily V contains five maize GATA and three *Arabidopsis* GATA, subfamily VI contains three maize GATA and seven *Arabidopsis* GATA, and subfamily has the same number of maize GATA and *Arabidopsis* GATA, five. It is worth noting that subfamily III contains much more maize genes than *Arabidopsis thaliana*. It is speculated that GATA in this subfamily formed multiple copies through gene replication during maize evolution, which may play an



**Fig. 1:** Phylogenetic evolution of GATA genes in maize and *Arabidopsis thaliana*



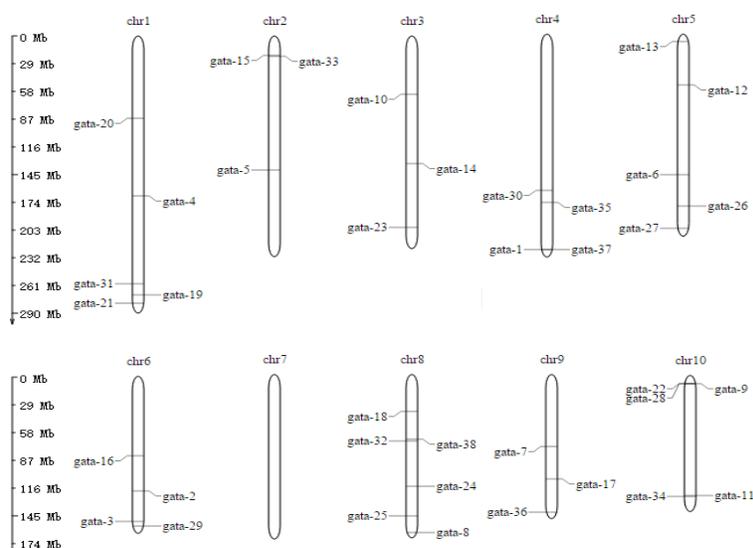
**Fig. 2:** Conserved domains of GATA amino acid sequence in maize

important role in maize development.

**Analysis of conserved domain of GATA transcription factor family protein sequence**

Conservative domain is a highly conservative domain in the process of biological evolution. The analysis of conserved

domain of amino acid sequence of GATA protein in maize showed that all 38 GATA transcription factors in maize contained GATA structure. Among them, 33 GATA transcription factors in maize contained only one structure of GATA (Fig. 2), such as GRMZM2G118214, GRMZM2G324131, GRMZM2G464037, GRMZM2G009530, GRMZM2G0671. These transcription



**Fig. 3:** Location of GATA transcription factor family genes on different chromosomes of maize

factors contain two GATA structures, and their location and number are diverse. It is noteworthy that three GATA proteins GATA-16 (GRMZM2G080509), GATA-13 (GRMZM2G065896), GATA-12 (GRMZM2G058479) contain GATA, tify, CCT and CCT structures, GATA-34 (GRMZM2G421212) contains GATA and ASXH structures, and GATA-31 (GRMZM2G065896), respectively. 396451) contains two Thioredoxin\_8 structures, Gin-synt\_C, MDD\_C, Pkinase\_Tyr and GATA. The conserved domain analysis indicated that the GATA transcription factors might have functional specificity through the increase of different domains.

### Location analysis of GATA gene on chromosome

According to the information of maize genome B73\_ref\_v2, it was determined that distribution of 38 GATA transcription factor family genes on 10 maize chromosomes (Fig. 3). The five transcription factors GATA-4, GATA-19, GATA-20, GATA-21 and GATA-31 were located on the first chromosome. GATA-5, GATA-15 and GATA-33 were the three transcription factor loci. On the second chromosome, GATA-10, GATA-14, GATA-23 are located on the third chromosome, GATA-1, GATA-30, GATA-35, GATA-37 on the fourth chromosome, GATA-6, GATA-12, GATA-13, GATA-26, GATA-27 on the fifth chromosome, GATA-2, G-37 on the fifth chromosome. ATA-3, GATA-16 and GATA-29 are located on chromosome 6. GATA-8, GATA-18, GATA-24, GATA-25, GATA-32 and GATA-38 on chromosome 8. GATA-7, GATA-17 and GATA-36 on chromosome 9. GATA-9, GATA-11, GATA-22, GATA-28 and GATA-36 on chromosome 9. The five transcription factors GATA-34 are located on chromosome 10, of which the seventh chromosome has no GATA transcription factor family gene. These results indicated that the distribution of GATA gene in maize chromosomes was dispersed and there was no centralized region.

### Physicochemical properties of GATA transcription factor family proteins

The amino acid composition and physicochemical properties of different GATA transcription factor family proteins are different, and the amino acid number, molecular weight, theoretical isoelectric point, fat coefficient, and hydrophobic average coefficient are different in different GATA transcription factors. The GATA-10 (GRMZM2G050616), amino acid number 152, molecular weight 16441.52, theoretical isoelectric point 9.89, fat coefficient 57.17, hydrophobicity-0.695. GATA-31 (GRMZM2G396451) has 1485 amino acids, 163386.21 molecular weight, 9.13 theoretical isoelectric point, 82.20 fat coefficient and -0.283 hydrophobicity (Table 1). Overall, the largest number of amino acids was GATA-31 (1485), the smallest number were GATA-10 (152) with smallest fat coefficient GATA-17 (43.89), and the greatest fat coefficient of GATA-31 (82.20). The GATA transcription factor family proteins in maize were rich in basic and acidic amino acids, of which 29 were alkaline and 9.20 were acidic. One is acidic, the most alkaline is GATA-37, its theoretical isoelectric point is 11.75, the most acidic is GATA-16, its isoelectric point is 5.16. The average hydrophobic coefficients of GATA transcription factor family proteins in maize were all negative, indicating that these GATA transcription factors were mainly hydrophobic proteins. Fat coefficient can be used as an index of protein stability. The higher the fat coefficient, the higher the protein stability. The fat coefficient of GATA transcription factors in maize is high, so it is known that GATA transcription factor family proteins are more stable.

### Subcellular localization

An online tool (<http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/>) was used to predict subcellular localization (Zhu et

**Table 1:** Physicochemical properties of GATA transcription factor family proteins

Gene	Locus	Amino acid number	Molecular weight	Theoretical isoelectric point	Fat coefficient	Average coefficient of hydrophobicity	Predicted location
GATA-1	AC184831	302	2.553184	9.86	65.30	-0.499	Nucleus.
GATA-2	AC194965	164	17607.89	9.56	58.48	-0.492	Nucleus.
GATA-3	AC202864	403	41685.40	5.56	71.36	-0.322	Nucleus.
GATA-4	GRMZM2G009530	306	34147.53	8.99	45.62	-0.910	Nucleus.
GATA-5	GRMZM2G025002	481	49438.24	6.51	53.14	-0.474	Nucleus.
GATA-6	GRMZM2G031983	395	42111.63	9.42	60.18	-0.440	Nucleus.
GATA-7	GRMZM2G039586	462	49012.51	8.88	67.32	-0.298	Nucleus.
GATA-8	GRMZM2G044576	400	42005.19	5.70	63.30	-0.356	Nucleus.
GATA-9	GRMZM2G048850	266	28076.59	7.81	53.76	-0.605	Nucleus.
GATA-10	GRMZM2G052616	152	16441.52	9.89	57.17	-0.695	Nucleus.
GATA-11	GRMZM2G054615	463	48766.84	9.14	57.49	-0.549	Nucleus.
GATA-12	GRMZM2G058479	382	40600.83	5.23	57.15	-0.660	Nucleus.
GATA-13	GRMZM2G065896	303	32369.68	8.76	63.60	-0.463	Nucleus.
GATA-14	GRMZM2G067171	486	51976.12	10.05	60.12	-0.724	Nucleus.
GATA-15	GRMZM2G077002	441	47780.70	8.71	59.12	-0.680	Nucleus.
GATA-16	GRMZM2G080509	382	40628.16	5.16	65.29	-0.565	Nucleus.
GATA-17	GRMZM2G101058	398	41151.58	8.52	43.89	-0.524	Nucleus.
GATA-18	GRMZM2G104390	164	17464.74	9.90	57.87	-0.605	Nucleus.
GATA-19	GRMZM2G110295	396	40896.09	7.87	49.04	-0.457	Nucleus.
GATA-20	GRMZM2G113098	385	39950.03	7.87	45.12	-0.534	Nucleus.
GATA-21	GRMZM2G114775	188	20588.33	10.02	52.61	-0.845	Nucleus.
GATA-22	GRMZM2G118214	486	51964.04	9.49	62.70	-0.648	Nucleus.
GATA-23	GRMZM2G123909	258	26165.26	7.98	53.68	-0.179	Nucleus.
GATA-24	GRMZM2G135381	409	42935.05	5.73	68.36	-0.469	Nucleus.
GATA-25	GRMZM2G138967	220	23837.84	9.76	52.59	-0.809	Nucleus.
GATA-26	GRMZM2G140669	419	44524.67	8.58	70.41	-0.403	Nucleus.
GATA-27	GRMZM2G163200	447	46031.20	5.66	62.42	-0.333	Nucleus.
GATA-28	GRMZM2G324131	696	76385.90	7.79	64.84	-0.726	Nucleus.
GATA-29	GRMZM2G325850	309	32552.44	8.67	51.10	-0.369	Nucleus.
GATA-30	GRMZM2G379005	430	44878.06	5.43	61.23	-0.433	Nucleus.
GATA-31	GRMZM2G396451	1485	163386.21	9.13	82.20	-0.283	Nucleus.
GATA-32	GRMZM2G397616	268	27782.01	8.75	48.36	-0.333	Nucleus.
GATA-33	GRMZM2G404973	487	52871.66	9.03	60.37	-0.629	Nucleus.
GATA-34	GRMZM2G421212	568	62500.53	6.96	67.50	-0.606	Nucleus.
GATA-35	GRMZM2G464037	497	51810.82	9.34	57.89	-0.579	Nucleus.
GATA-36	GRMZM2G532534	348	37892.38	8.64	52.33	-0.533	Nucleus.
GATA-37	GRMZM5G879778	253	27037.71	11.75	56.72	-0.654	Nucleus.
GATA-38	GRMZM5G887975	180	19527.96	9.97	51.17	-0.654	Nucleus.

al. 2012). The results showed that all 38 GATA transcription factor families are located in the nucleus, indicating that their role in regulating transcription (Table 1).

### Analysis of GATA expression in maize

Using transcriptome sequencing data published by Stelpflug *et al.* (2016), we analyzed the tissue-specific expression of 38 GATA families in maize, and drew a heat map based on the FPKM value of each gene. The tissues analyzed included germinated seeds, different regions of roots, seedlings, stems at different locations, apical meristem of stems, leaves, internodes, spikes, anthers and maize whiskers (Reyes *et al.* 2004). The expression patterns of GATA transcription factor family genes are different in different tissues and developmental stages of maize (Fig. 4). GATA-1, GATA-23, GATA-29, GATA-5, GATA-32, GATA-35, GATA-38, GATA-37, GATA-9 and GATA-36 were not expressed in the tissues and developmental stages analyzed. GATA-26, GATA-31, GATA-21, GATA-27, GATA-30, GATA-12, GATA-16, GATA-11 and GATA-33 all showed

high levels of expression during the period under analysis. GATA-10, GATA-18, GATA-34, GATA-2, GATA-15, GATA-13, GATA-14 and GATA-22 were expressed at a moderate level in all stages and organizations during the analysis period. Other GATA genes have different expression patterns at different stages. For example, GATA-27 and GATA-3 are obviously expressed in Embryo. GATA-17 is highly expressed in coleoptile, seed and internode. GATA-6 and GATA-7 are specifically expressed in leaves during the analysis period. Specific expression was found in seeds and endosperm. The expression characteristics of GATA transcription factors indicated that the GATA family genes in maize were functionally differentiated.

### Discussion

Transcription factors play an important regulatory role in plant growth and development and response to environmental change. They are the key links in regulating various physiological activities. In recent years, the function and significance of GATA in non-maize have been reported



structures are not complicated. The fat coefficient of GATA transcription factors in maize is high, so it is known that GATA transcription factor family proteins are more stable.

In sub-cellular localization, 38 GATA transcription factor families were all located in the nucleus, indicating their role in regulating transcription. According to chromosome mapping, there are 6 GATA genes on the eighth chromosome of maize, which belong to the largest number. The seventh maize chromosome has no GATA transcription factor family gene.

The analysis of expression pattern of maize GATA transcription factor family showed a tissue-independent manner for most GATA genes, however, GATA-27, GATA-17, GATA-6, GATA-7 and GATA-28 were specifically expressed in several tissues. In *Arabidopsis*, GATA factors are found to be under the control of multiple developmental and environmental signals (Behringer and Schwechheimer 2015). The GATA transcription factors might play different roles in the process of maize growth.

## Conclusion

The GATA transcription factor family in maize was identified by bioinformatics and molecular biology methods, and its function in maize was preliminary analyzed. This experiment provides some basic data for further utilization of GATA transcription factors in maize breeding.

## Acknowledgements

The authors acknowledge the Scientific Research Project fund of Jilin Provincial Education Department (JJKH20190979KJ), Science and Technology Innovation Development Project of Jilin City (#201831781 to X. Y.), Natural Science Foundation of Jilin Province of China (#20180101233JC to Z.-Y.X.), and Doctor Talent Research Foundation of Jilin Agricultural Science and Technology University (#20185002 to L. J.).

## References

- Behringer C, C Schwechheimer (2015). B-GATA transcription factors—insights into their structure, regulation, and role in plant development. *Front Plant Sci* 2:1–12
- Chen GL, HH Zu, H Zhang, BP Xue, T Ren, C Zhou (2018). Bioinformatics analysis of the GATA transcription factor family in *Ziziphys jujube*. *Mol Plant Breed* 16:4863–4871
- Chen H, H Shao, K Li, D Zhang, S Fan, Y Li, M Han (2017). Genome-wide identification, evolution, and expression analysis of GATA transcription factors in apple. *Gene* 627:460–472
- Liu C, XX Li, YL Su, YF Guo (2015). Genome-wide identification, phylogenetic analysis and expression profiling of the SBP transcription factor family in *Nicotiana tobacum*. *Chin Tobac Sci* 36:1–11
- Reyes JC, MI Muro, FJ Florencio (2004). The GATA family of transcription factors in Arabidopsis and rice. *Plant Physiol* 134:120–121
- Shi YH, ZR Xu (2005). Research progress of GATA transcription factors. *Biol Bull* 25:1–2
- Stelpflug SC, RS Sekhon, B Vaillancourt (2016). An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genomics* 9:85–94
- Xie T, S Wang, L Huang, X Wang, LP Kang, LP Guo (2014). Transcriptome-based bioinformatics analysis of *Arnebia euchroma* ERF transcription factor family. *Chin J Trad Chin Med* 39:4732–4739
- Yuan Q, CL Zhang, TT Zhao, XY Xu (2017a). Bioinformatics analysis of GATA transcription factor in Pepper. *Chin Agric Sci Bull* 33:24–31
- Yuan Q, CL Zhang, TT Zhao, XY Xu (2017b). Research advances of GATA transcription factor in plant. *Mol Plant Breed* 15:1702–1707
- Zhao JL, WJ Yao, SJ Wang, YB Jiang, BR Zhou (2015). AP2/ERF gene family in *Populus trichocarpa* by bioinformatics. *J Northeast For Univ* 43:21–29
- Zhao JR, S Wang, M Li, HY Lv, DW Wang, YQ Ge, X Wei, WC Yang (2018). Current status and perspective of maize breeding. *J Plant Genet Res* 19:435–446
- Zhao JR, RH Wang, XX Liu (2016). The current situation of maize industry and the development trend of biological breeding in China. *Biotechnol Busin* 3:45–52
- Zhu J, Z Sun, D Shi, S Song, L Lian, A Ren, H Yu, M Zhao (2019). Dual functions of area, a GATA transcription factor, on influencing ganoderic acid biosynthesis in *Ganoderma lucidum*. *Environ Microbiol* 21:4166–4179
- Zhu S, L Zeng, PZ Wu, YP Chen, HW Jiang, GJ Wu, MR Li (2012). Cloning and sequence analysis of the cDNA of plastid-located glycerol-3-phosphate acyltransferase (JcGPAT2) gene from *Jatropha curcas* L. *Guangdong Agric Sci* 39:1–5