Data in Brief 20 (2018) 2027–2035

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Identification of putative flowering genes and transcription factors from flower *de novo* transcriptome dataset of tuberose (*Polianthes tuberosa* L.)

Jayanthi Madhavan [a,*], Pawan Jayaswal [b], Kanchan B.M. Singh [a], Uma Rao [a]

[a] *Division of Nematology, Indian Agricultural Research Institute, New Delhi, India*
[b] *National Research Centre for Plant Biotechnology, Pusa, New Delhi 110012, India*

ARTICLE INFO

Article history:
Received 6 June 2018
Received in revised form
11 September 2018
Accepted 19 September 2018
Available online 22 September 2018

Keywords:
Tuberose
Amaryllidaceae
Transcriptome analysis
KEGG
Flower specific genes
Transcription factors

ABSTRACT

Polianthes tuberosa is commercially popular because of their economic importance in floriculture for cut and loose flowers and in perfume industry because of the unique fragrance. Despite its commercial importance, no ready-to-use transcript sequence information is available in the public database. We have sequenced the RNA obtained from tuberose flowers using the Illumina HiSeq. 2000 platform and have carried out a *de novo* analysis of the transcriptome data. The *de novo* assembly generated 11,100 transcripts. These transcripts represent a total of 7876 unigenes that were considered for downstream analysis. These 7876 unigenes, which was further annotated using blast2go and KEGG pathways, were also assigned. Tuberose transcripts were also assigned to metabolic pathways using the Kyoto Encyclopedia of Genes and Genomes database to determine their biochemical functions. 4591 of the tuberose transcripts matched to genes in KEGG pathways and 66 transcripts were mapped to the Flavonoid biosynthesis pathway. 21 flowering genes have been identified in this tuberose transcriptome. Transcription factor analysis helped in the identification of a large number of transcripts similar to key genes in the flowering regulation network of *Arabidopsis thaliana*. Among the transcription factors identified “NAC” which is associated with plant stress response represented the most abundant category followed by APETALA2 (AP2)/ethylene-responsive element binding

[*] Corresponding author.
E-mail address: jayman21@gmail.com (J. Madhavan).

https://doi.org/10.1016/j.dib.2018.09.051

proteins (EREBPs) which plays various role in floral organ identity and respond to different biotic and abiotic stress.

Specifications table

Subject area	*Plant Biotechnology and Bioinformatics*
More specific subject area	*Transcriptome*
Type of data	*Table, text file, graph, figure*
How data was acquired	*Illumina Hiseq. 2000 platform at SciGenom Next-Gen sequencing facility*
Data format	*Analyzed*
Experimental factors	*RNA was isolated from flowers of Polianthes tuberosa*
Experimental features	*Transcriptome sequence of tuberose flower and de novo analysis for identification of flowering genes and transcription factors*
Data source location	*New Delhi, India*
Data accessibility	*Data is with this article and the raw sequence data generated has been deposited in the SRA database (*http://www.ncbi.nlm.nih.gov/bioproject/321962*) for public access (BioSample accession ID: SAMN05006898).*

Value of the data

- This is the first report of *de novo* transcriptome analysis of *Polianthes tuberosa* flower. Tuberose transcripts were assigned KEGG pathways from the transcriptome data. Flowering genes and transcription factors were identified from the transcriptome data successfully.
- Transcriptome data will provide a strong foundation for research on gene expression, genomics and functional genomics in *Polianthes tuberosa* and other important members of Amaryllidaceae.
- The data generated during this work has not only added so much of information on a plant which had no genomic information on the public domain but also shall help in the studies of other economically important plants like daffodils, snowflakes, onions and garlic belonging to the same family.
- The data will help in the better understanding of expression patterns and their relation to function and regulation, and also the genetic mechanisms, evolutionary relationships between tuberose and other plants.
- This transcriptomic analysis has opened up the prospects for a better understanding of its genomics and we have updated the current gene resource.

1. Data

In spite of its considerable industrial importance, genomic information on tuberose is very scarce. There are no public Expressed Sequence Tags (EST) or ready-to-use transcripts for *Polianthes tuberosa*. This is for the first time a high-throughput, RNA sequencing (RNA-Seq) of the *P. tuberosa* flower transcriptome was carried out to generate a database that will be useful for further functional analyses. An overview of the sequencing assembly of *P. tuberosa* transcriptome data is presented in Table 1. The length distribution of unigenes is shown in the Fig. 1. The blast result showed that unigenes returned 79.76% (6282) significant hits against the reported datasets. When considering the

Table 1
Summary of transcriptome sequence assembly of *Polianthes tuberosa* data.

Content	Contig	Unigene
Number	11,100	7876
L50	2692	2000
Minimum length	52	52
N80	511	558
N50	968	1010
N20	1677	1705
Maximum length	9548	9548
Total number of bases	8,238,911	6,236,175

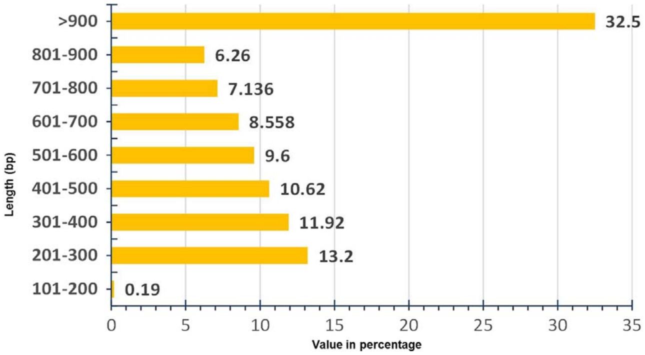


Fig. 1. Length distribution of 7876 Unigene sequences.

annotation by species, significant similarity to *Elaeis guineensis* followed by *Phoenix dactylifera* both belonging to the monocotyledons was obtained (Fig. 2).

Using gene ontology, 1446 ESTs were classified to cellular component category, 2521 ESTs were classified for biological process and 1493 ESTs were classified under molecular function category. A summary with the number and percentage of unigenes annotated in each GO slim term is shown (Fig. 3). According to the data 4122 unique sequences were classified into 24 COG categories (Fig. 4). KEGG Orthology (KO identifiers) for the unigenes were retrieved (Supplementary Table S1a; Fig. 5). As

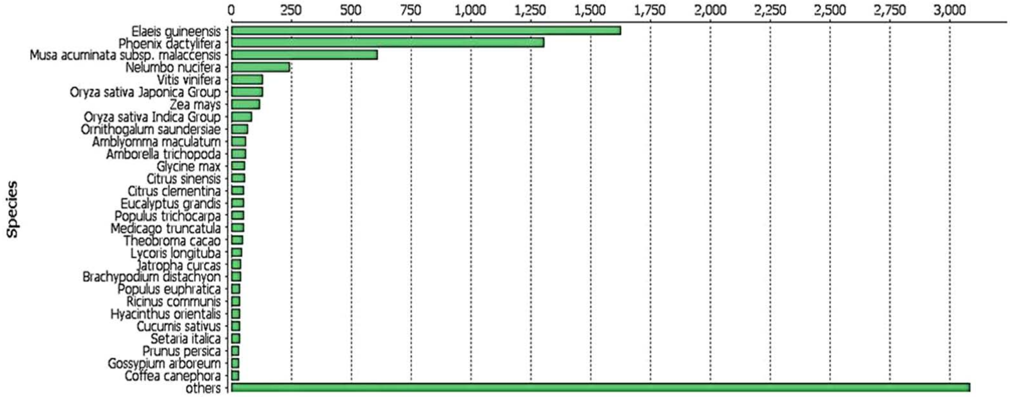


Fig. 2. Top BLAST hit species distribution, obtained by BLASTx against the NCBI non-redundant (nr) protein database. The number of top BLAST hits per species is shown on the *x*-axis. Only the 29 most represented species are shown. The complete number of top hits of other organisms is 3080.

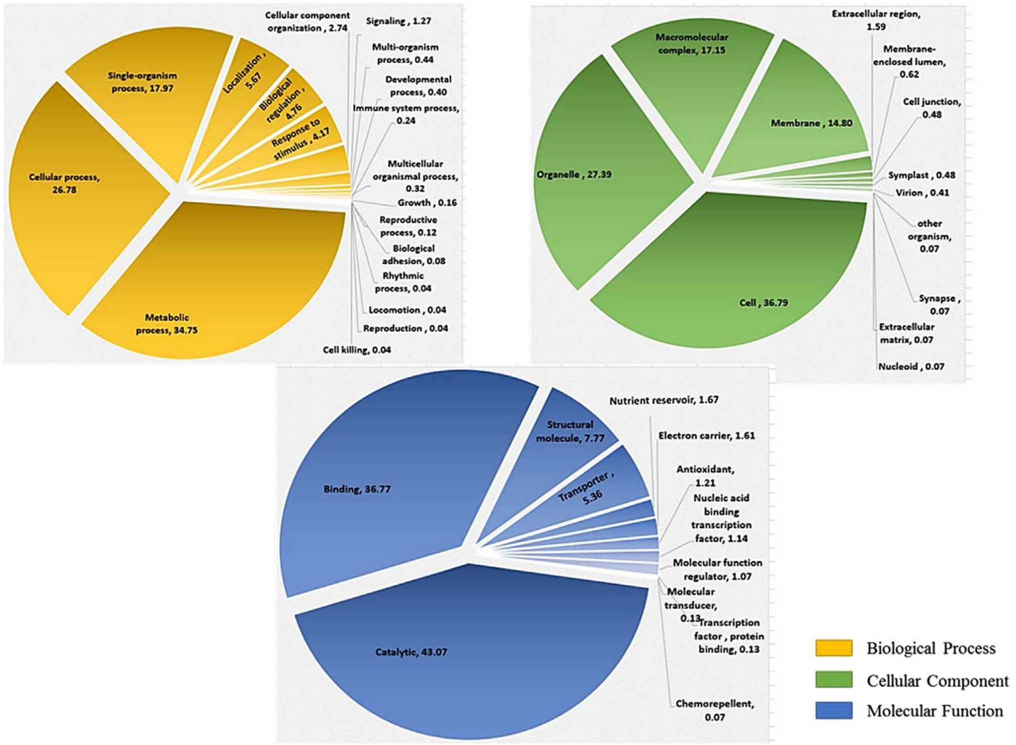


Fig. 3. Gene ontology annotations of the 7876 contigs of the *Polianthes tuberosa* transcriptome dataset into three different subcategories like biological process, cellular component and molecular function. Mentioned percentage value indicates the protein-coding *Polianthes tuberosa* transcript assigned to each category.

many as 4591 of the tuberose transcripts matched to genes in KEGG pathways (Supplementary Table S1b). We have identified 21 unigenes which showed homology to *Arabidopsis thaliana* flowering genes (Table 2). Analysis of transcription factor in tuberose revealed a total of 511 unigenes, representing 6.48% of the transcriptome classified into 59 putative transcription factors (TF) families (Supplementary Table S2; Fig. 6).

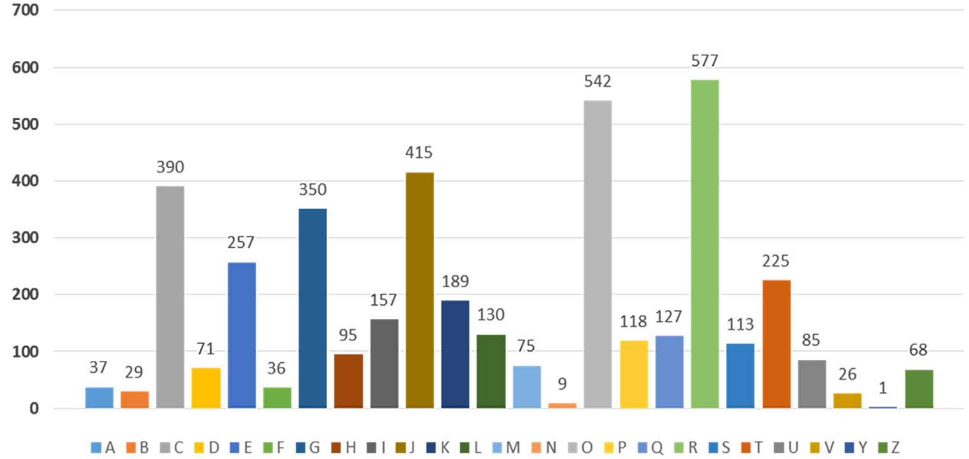
2. Experimental design, materials, and methods

2.1. Plant material

Fully opened tuberose flowers of cultivar Shringar were collected and were immediately frozen in liquid nitrogen and stored at −80 °C.

2.2. RNA extraction, cDNA library construction and sequencing

Total RNA was extracted from frozen flower tissues using 596 Nucleospin RNA isolation kit (Macherey-Nagel GmbH & Co. KG, Duren, Germany). Agilent 2100 Bioanalyzer (Agilent Technologies) was used to assess the quality and quantity of RNA. RNA with an RNA integrity number (RIN) of 8.0 was only considered mRNA purification. OligodT beads (Illumina® TruSeq® RNA Sample Preparation Kit v2) were used to purify mRNA from one microgram of total RNA. Elevated temperature (90 °C) in presence of divalent cations was used to achieve the fragmentation of the purified mRNA. cDNA synthesis was done using random hexamers with Superscript II Reverse Transcriptase



A	RNA processing and modification	M	Cell wall/membrane/envelope biogenesis
B	Chromatin structure and dynamics	N	Cell motility
C	Energy production and conversion	O	Posttranslational modification, protein turnover, chaperones
D	Cell cycle control, cell division, chromosome partitioning	P	Inorganic ion transport and metabolism
E	Amino acid transport and metabolism	Q	Secondary metabolites biosynthesis, transport and catabolism
F	Nucleotide transport and metabolism	R	General function prediction only
G	Carbohydrate transport and metabolism	S	Function unknown
H	Coenzyme transport and metabolism	T	Signal transduction mechanisms
I	Lipid transport and metabolism	U	Intracellular trafficking, secretion, and vesicular transport
J	Translation, ribosomal structure and biogenesis	V	Defense mechanisms
K	Transcription	Y	Nuclear structure
L	Replication, recombination and repair	Z	Cytoskeleton

Fig. 4. Distribution of clusters of orthologous groups (COGs) of 4122 unigene sequences into 24 different groups.

(Invitrogen Life Technologies). Agencourt Ampure XP SPRI beads (Beckman-Coulter) were used to clean the cDNA. Illumina adapters were ligated to the cDNA molecules after end repair and the addition of an ‘A’ base followed by SPRI clean-up. The resultant cDNA library was amplified using PCR for the enrichment of adapter-ligated fragments, quantified using a Nanodrop spectrophotometer (Thermo Scientific) and validated for quality with a Bioanalyzer (Agilent Technologies). The libraries were then sequenced on Illumina Hiseq. 2000 platform at SciGenom Next-Gen sequencing facility, Cochin, India.

2.3. *Sequence data assembly and analysis*

NGSQC Toolkit version v2.3.3 [1] was used to remove low quality reads (Phred score $<$ 30) and to generate sequencing statistics. High quality paired end filtered reads (15.9 gb) obtained were used for *de-novo* assembly using Velvet (v.1.2.08) and Oases (v.0.2.08) pipeline [2]. Velveth assembly was done with various k-mer range (71- 83) and optimal assembly was attained at k-mer 83. Oases tool was used to identify non-overlapping isoforms/splice variants at minimum transcript length 100. Since our initial target was to identify unique genes. Thus, transcripts were subjected for clustering using CD-HIT-EST [3] 90% similarity. ORF Predictor web server (http://bioinformatics.ysu.edu/tools/OrfPredictor.html) [4] was used to predict proteins from the all non-redundant transcripts ($\geq$ 100 bp) using the default cut-off value of $1e-5$, and 7876 proteins were predicted which were considered for the annotation. The raw sequence data generated has been deposited in the SRA database (http://www.ncbi.nlm.nih.gov/bioproject/321962) for public access (BioSample accession ID: SAMN05006898).

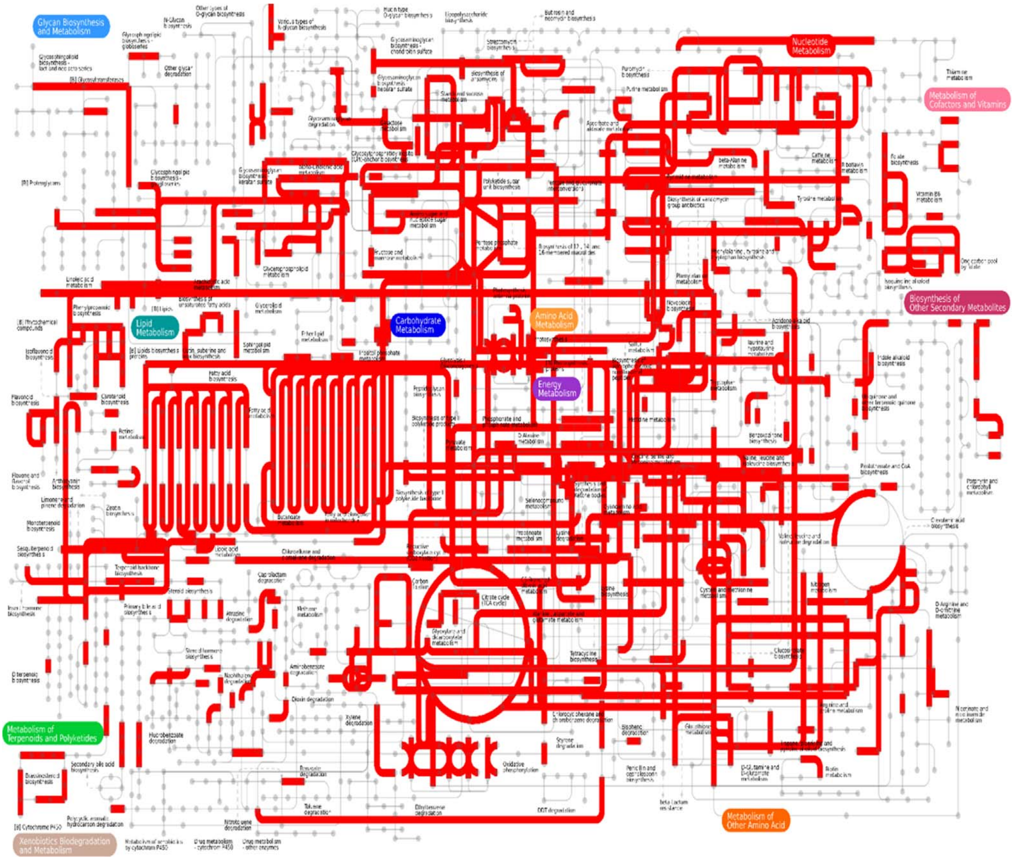

Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid

Fig. 5. Metabolic pathways active in tuberose as revealed by the transcriptomic analysis using iPATH2 interactive pathway explorer.

2.4. Functional annotation and biological classification of transcripts

Functional annotation of predicted tuberose transcripts was performed using blast2go pipeline on default settings [5]. BLASTP [6] were performed with an *E*-value of 1e − 5 to align against NCBI non-redundant (nr) protein database for homology search. Blast results (xml format) were imported to Blast2GO V.3.0.11. GO annotations were performed with default settings and following GO annotation, an Interproscan [7] was performed and results were merged to the GO annotations.

2.5. Identification of flowering genes

Homologous flowering gene in tuberose plant were identified using BLASTN programme 306 gene of *A. thaliana* (http://www.phytosystems.ulg.ac.be/florid/) database.

2.6. Identification of transcription factors

For the identification of transcription factor in tuberose plant data we used PlnTFDB (3.0) database (http://plntfdb.bio.uni-potsdam.de/v3.0/). Standalone BLASTN programme used for the identification of homologous TF in tuberose plant and output has parsed from BLAST Parser v1.2.6.14 programme (http://geneproject.altervista.org/) and filtered with 60% identity and 100 bit score.

Table 2
List of flowering genes homologous to *Arabidopsis thaliana*.

S.no	Tuberose	Flowering gene	Accession No.	Functions	References
1	TUBEROSE_186	NM_114279.4 Ath DNAJ homologue 3 (J3), mRNA	AT3G44110	Flowering promoter; mediates the transcriptional regulation of two floral pathway integrators, FLOWRING LOCUS T and SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 and regulates flowering time in *Arabidopsis thaliana*	[8,9]
2	TUBEROSE_203	NM_118595.5 Ath phosphoglucose isomerase 1 (PGI1), mRNA	AT4G24620	Carbohydrate metabolism, important role in floral initiation, flowering delayed in mutants	[10]
3	TUBEROSE_316	NM_001333000.1 Ath WWE protein-protein interaction domain protein family (RCD1), mRNA	AT1G32230	RCD1–6 mutant showed reduced flowering	[11]
4	TUBEROSE_317	NM_001333000.1 Ath WWE protein-protein interaction domain protein family (RCD1), mRNA	AT1G32230	RCD1–6 mutant showed reduced flowering	[11]
5	TUBEROSE_370	NM_125149.3 Ath CONSTANS-like 5 (COL5), mRNA	AT5G57660	Induce flowering in short day *Arabidopsis thaliana*	[12]
6	TUBEROSE_385	NM_127738.5 Ath cold, circadian rhythm, and RNA binding 2 (GRP7), mRNA	AT2G21660	Promotes floral transition partly by down regulating FLC	[13]
7	TUBEROSE_430	NM_111158.4 Ath GAST1 protein homolog 5 (GASA5), mRNA	AT3G02885	GASA5 is a negative regulator of GA-induced flowering	[14]
8	TUBEROSE_433	NM_001342189.1 Ath homeobox protein ATH1 (ATH1), mRNA	AT4G32980	ATH1 regulates FLC	[15]
9	TUBEROSE_515	NM_130127.2 Ath AGAMOUS-like 6 (AGL6), mRNA	AT2G45650	AGL6 acts as a floral promoter with a dual role, the inhibition of the transcription of the FLC/MAF genes and the promotion of FT expression in Arabidopsis	[16]
10	TUBEROSE_521	NM_001035973.3 AthTransducin family protein / WD-40 repeat family protein (TPL), mRNA	AT1G15750	Represses flowering in *Arabidopsis thaliana*	[17,18]
11	TUBEROSE_532	NM_001337962.1 Ath ubiquitin-specific protease 13 (UBP13), mRNA	AT3G11910	Control of the circadian clock and photoperiodic flowering	[19]
12	TUBEROSE_589	NM_125149.3 Ath CONSTANS-like 5 (COL5), mRNA	AT5G57660	Induce flowering in short day *Arabidopsis thaliana*	[12]
13	TUBEROSE_589	NM_125149.3 Ath CONSTANS-like 5 (COL5), mRNA	AT5G57660	Induce flowering in short day *Arabidopsis thaliana*	[12]
14	TUBEROSE_597	NM_001344334.1 Ath RNA-binding (RRM/RBD/RNP motifs) family protein mRNA	AT5G40490	HLP1 regulates flowering by alternative polyadenylation	[20]
15	TUBEROSE_645	NM_001332707.1 Athcryptochrome-interacting basic-helix-loop-helix 5 (CIB5), mRNA	AT1G26260	Regulates flowering time redundantly with CIB1.	[21]
16	TUBEROSE_685	NM_102124.3 Ath gigantea protein (GI), mRNA	AT1G22770	promotes flowering under long days in a circadian clock-controlled flowering pathway	[22]
17	TUBEROSE_698	NM_128569.4 Ath UDP-Glycosyltransferase superfamily protein (UGT87A2), mRNA	AT2G30140	Regulates flowering time via the flowering repressor FLC	[23]
18	TUBEROSE_740	NM_114187.5 Ath sucrose synthase 4 (SUS4), mRNA	AT3G43190	Promotes flowering	[24]
19	TUBEROSE_770	NM_101307.5 Ath ubiquitin carrier protein 1 (UBC1), mRNA	AT1G14400	Monoubiquitination of H2B via UBC1 regulates flowering time	[25,26]
20	TUBEROSE_783	NM_125119.4 Ath Galactose oxidase/kelch repeat superfamily protein (ZTL), mRNA	AT5G57360	Control of flowering time	[27]

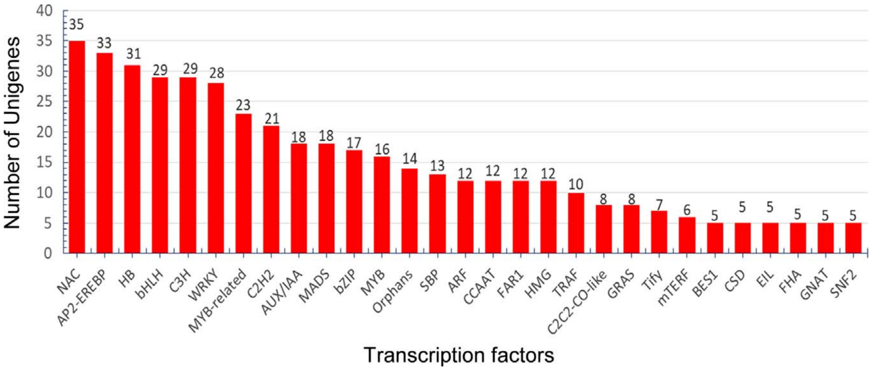


Fig. 6. Transcription factor in tuberose distribution of 442 copies (≥ 5) of TF distributed among 29 different large categories.

Acknowledgments

The authors gratefully acknowledge Director and Joint Director, Indian Agricultural Research Institute, New Delhi (India) for providing the research facilities and Department of Science and Technology, India (SERB grant EMR/2015/001647) for funding the research.

Transparency document. Supplementary material

Transparency document associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.09.051.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.09.051.

References

[1] R.K. Patel, M. Jain, NGS QC toolkit: a toolkit for quality control of next generation sequencing data, PLoS One 7 (2007) e30619. https://doi.org/10.1371/journal.pone.0030619.

[2] D.R. Zerbino, E. Birney, Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829. https://doi.org/10.1101/gr.074492.107.

[3] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases, Bioinformatics 17 (2001) 282–283.

[4] X.J. Min, G. Butler, R. Storms, A. Tsang, OrfPredictor: predicting protein-coding regions in EST-derived sequences, Nucleic Acids Res. 33 (2005) W677–W680. https://doi.org/10.1093/nar/gki394.

[5] A. Conesa, S. Götz, Blast2GO: a comprehensive suite for functional analysis in plant genomics, Int. J. Plant Genom. 2008 (2008) 619832. https://doi.org/10.1155/2008/619832.

[6] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

[7] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, et al., The InterPro database, an integrated documentation resource for protein families, domains and functional sites, Nucleic Acids Res. 29 (2001) 37–40.

[8] L. Shen, H. Yu, J3 regulation of flowering time is mainly contributed by its activity in leaves, Plant Signal. Behav. 6 (2011) 601–603. https://doi.org/10.4161/psb.6.4.15375.

[9] L. Shen, Y.G.G. Kang, L. Liu, H. Yu, The J-domain protein J3 mediates the integration of flowering signals in *Arabidopsis*, Plant Cell 23 (2011) 499–514. https://doi.org/10.1105/tpc.111.083048.

[10] T.S. Yu, W.L. Lue, S.M. Wang, J. Chen, Mutation of *Arabidopsis* plastid phosphoglucose isomerase affects leaf starch synthesis and floral initiation, Plant Physiol. 123 (2000) 319–326. https://doi.org/10.1104/pp.123.1.319.

[11] S. Teotia, R.S. Lamb, The paralogous genes RADICAL-INDUCED CELL DEATH1 and SIMILAR TO RCD ONE1 have partially redundant functions during *Arabidopsis* development, Plant Physiol. 151 (2009) 180–198. https://doi.org/10.1104/pp.109.142786.

[12] M. Hassidim, Y. Harir, E. Yakir, et al., Over-expression of CONSTANS-LIKE 5 can induce flowering in short-day grown *Arabidopsis*, Planta 230 (2009) 481–491. https://doi.org/10.1007/s00425-009-0958-7.

[13] C. Streitner, S. Danisman, F. Wehrle, et al., The small glycine-rich RNA binding protein AtGRP7 promotes floral transition in *Arabidopsis thaliana*, Plant J. 56 (2008) 239–250. https://doi.org/10.1111/j.1365-313X.2008.03591.x.

[14] S. Zhang, C. Yang, J. Peng, et al., *GASA5*, a regulator of flowering time and stem growth in *Arabidopsis thaliana*, Plant Mol. Biol. 69 (2009) 745. https://doi.org/10.1007/s11103-009-9452-7.

[15] M. Proveniers, B. Rutjens, M. Brand, S. Smeekens, The *Arabidopsis* TALE homeobox gene *ATH1* controls floral competency through positive regulation of FLC, Plant J. 52 (2007) 899–913. https://doi.org/10.1111/j.1365-313X.2007.03285.x.

[16] S.K. Yoo, X. Wu, J.S. Lee, J.H. Ahn, AGAMOUS-LIKE 6 is a floral promoter that negatively regulates the FLC/MAF clade genes and positively regulates FT in *Arabidopsis*, Plant J. 65 (2011) 62–76. https://doi.org/10.1111/j.1365-313X.2010.04402.x.

[17] M. Graeff, D. Straub, T. Eguen, U. Dolde, V. Rodrigues, R. Brandt, et al., Microprotein-mediated recruitment of CONSTANS into a TOPLESS trimeric complex represses flowering in *Arabidopsis*, PLoS Genet. 12 (2016) e1005959. https://doi.org/10.1371/journal.pgen.1005959.

[18] G.S. Goralogia, T.K. Liu, L. Zhao, P.M. Panipinto, et al., CYCLING DOF FACTOR 1 represses transcription through the TOPLESS co-repressor to control photoperiodic flowering in *Arabidopsis*, Plant J. 92 (2017) 244–262. https://doi.org/10.1111/tpj.13649.

[19] X. Cui, F. Lu, Y. Li, Y. Xue, Y. Kang, et al., Ubiquitin-specific proteases UBP12 and UBP13 act in circadian clock and photoperiodic flowering regulation in *Arabidopsis*, Plant Physiol. 162 (2013) 897–906. https://doi.org/10.1104/pp.112.213009.

[20] Y. Zhang, L. Gu, Y. Hou, L. Wang, X. Deng, et al., Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation, Cell Res. 25 (2015) 864–876. https://doi.org/10.1038/cr.2015.77.

[21] H. Liu, X. Yu, K. Li, et al., Photoexcited CRY2 interacts with CIB1 to regulate transcription and floral initiation in *Arabidopsis*, Science 322 (2008) 1535–1539. https://doi.org/10.1126/science.1163927.

[22] H.J. Park, W.Y. Kim, D.J. Yun, A role for GIGANTEA: keeping the balance between flowering and salinity stress tolerance, Plant Signal. Behav. 8 (2013) e24820. https://doi.org/10.4161/psb.24820.

[23] B. Wang, S.H. Jin, H.Q. Hu, Y.G. Sun, et al., UGT87A2, an *Arabidopsis* glycosyltransferase, regulates flowering time via FLOWERING LOCUS C, New Phytol. 194 (2012) 666–675. https://doi.org/10.1111/j.1469-8137.2012.04107.x.

[24] P.J. Seo, J. Ryu, S.K. Kang, C.M. Park, Modulation of sugar metabolism by an INDETERMINATE DOMAIN transcription factor contributes to photoperiodic flowering in *Arabidopsis*, Plant J. 65 (2011) 418–429. https://doi.org/10.1111/j.1365-313X.2010.04432.x.

[25] Y. Cao, Y. Dai, S. Cui, L. Ma, Histone H2B monoubiquitination in the chromatin of FLOWERING LOCUS C regulates flowering time in *Arabidopsis*, Plant Cell 20 (2008) 2586–2602. https://doi.org/10.1105/tpc.108.062760.

[26] L. Xu, R. Ménard, A. Berr, J. Fuchs, V. Cognat, et al., The E2 ubiquitin-conjugating enzymes, AtUBC1 and AtUBC2, play redundant roles and are involved in activation of *FLC* expression and repression of flowering in *Arabidopsis thaliana*, Plant J. 57 (2009) 279–288. https://doi.org/10.1111/j.1365-313X.2008.03684.x.

[27] Y.H. Song, D.A. Estrada, R.S. Johnson, et al., Distinct roles of FKF1, GIGANTEA, and ZEITLUPE proteins in the regulation of CONSTANS stability in *Arabidopsis* photoperiodic flowering, Proc. Natl. Acad. Sci. USA 111 (2014) 17672–17677. https://doi.org/10.1073/pnas.1415375111.