ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 18(1): 1-13 (2022)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

BIOMEDICAL INFORMATICS

www.bioinformation.net

Volume 18(1)

Research Article

Received October 21, 2021; Revised October 31, 2021; Accepted October 31, 2021, Published January 31, 2022

DOI: 10.6026/97320630018001

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article. The authors declare that a draft version of this article is made open access at https://www.biorxiv.org/content/10.1101/2021.06.23.449592v1

Declaration on official E-mail:
The corresponding author declares that official e-mail from their institution is not available for all authors

Comments from readers:
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

Edited by P. Kangueane
Citation: Ambardar ***et al.*** Bioinformation 18(1): 1-13 (2022)

Insights from the analysis of draft genome sequence of *Crocus sativus* L.

Sheetal Ambardar[1], Jyoti Vakhlu[2] & Ramanathan Sowdhamini[1,3,*]

[1]National Center for Biological Sciences, Bellary Road, Bengaluru 560065; [2]School of Biotechnology, University of Jammu, J&K, India, 180006; [3]Institute of Bioinformatics and Applied Biotechnology, Bengaluru 560100, India; *Corresponding author: Ramanathan Sowdhamini - E-mail: mini@ncbs.res.in

Abstract:
Saffron (*Crocus sativus* L.) is the low yielding plant of medicinal and economic importance. Therefore, it is of interest to report the draft genome sequence of *C. sativus*. The draft genome of *C. sativus* has been assembled using Illumina sequencing and is 3.01 Gb long covering 84.24% of genome. *C. sativus* genome annotation identified 53,546 functional genes (including 5726 transcription factors), 862,275 repeats and 964,231 SSR markers. The genes involved in the apocarotenoids biosynthesis pathway (crocin, crocetin, picrocrocin, and safranal) were found in the draft genome analysis.

Keywords: *Crocus sativus*, *de-novo* genome assembly, apocarotene biosynthesis pathway, MYB TFs, SSR markers, Orthology analysis.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 1-13 (2022)

Background:
Plant genomics, with the increasing number of whole genome sequences available, has unlocked the genetic treasures that would be impossible in absense of the genome sequence. Though second and third generation sequencing technologies, coupled with ever advancing bioinformatic tools/pipelines, have made the sequencing of complex and huge genomes economical and easy, but till date there are only approximately 1886 plant genome sequences available in databanks (NCBI: https://www.ncbi.nlm.nih.gov/assembly). Some of the recently sequenced and assembled plant genomes are rice [1], maize [2], asparagus [3], wheat [4] and tea [5] etc., however the genome of the plants belonging to Crocus genus or *Iridaceae* family, have not been reported so far. Saffron (*C. sativus*) referred as 'Golden Condiment' is world's most expensive spice costing about 70,000 INR/pound, with medicinal properties and cosmetic uses [6]. More than 150 volatile and aroma-yielding compounds contribute to the flavor, color, and aroma of the saffron spice, wherein the main chemical constituents in the stigma of saffron are crocin, crocetin, picrocrocin, and safranal [7]. *C. sativus* is an autumn-flowering perennial sterile triploid plant (2n = 24) with, ~3.5 Gb haploid genome [8,9]. Being sterile, it fails to produce viable seeds and reproduces vegetatively by underground corms and is reported to lack genetic variation. Various molecular markers (RAPD, ISSR, AFLP, SSR microsatellites) and epigenetic approaches have suggested the existence of limited genetic variability [10 - 13]. To discover authentic genetic markers, mining genes for secondary metabolites and improvement of breeding, sequencing of its genome was the only alternative. In addition, it's ancestry is also controversial that could be also settled, if its complete genome sequence is available [14, 15]. Hybrid sequencing approaches, comprising of second and third generation sequencing technologies, have facilitated sequencing of complex genomes economically. Illumina sequencing technology is preferred in combination of other sequencing technologies for first sequencing attempt, as it generates good sequencing data for better genome coverage and has low error rate as compared to third generation sequencing technologies [16]. Therefore, it is of interest to document data to gain insights from the preliminary analysis of draft genome sequence of *Crocus sativus* L. It should be noted that a draft version of this article has been made open access at the Biorxiv repository [17].

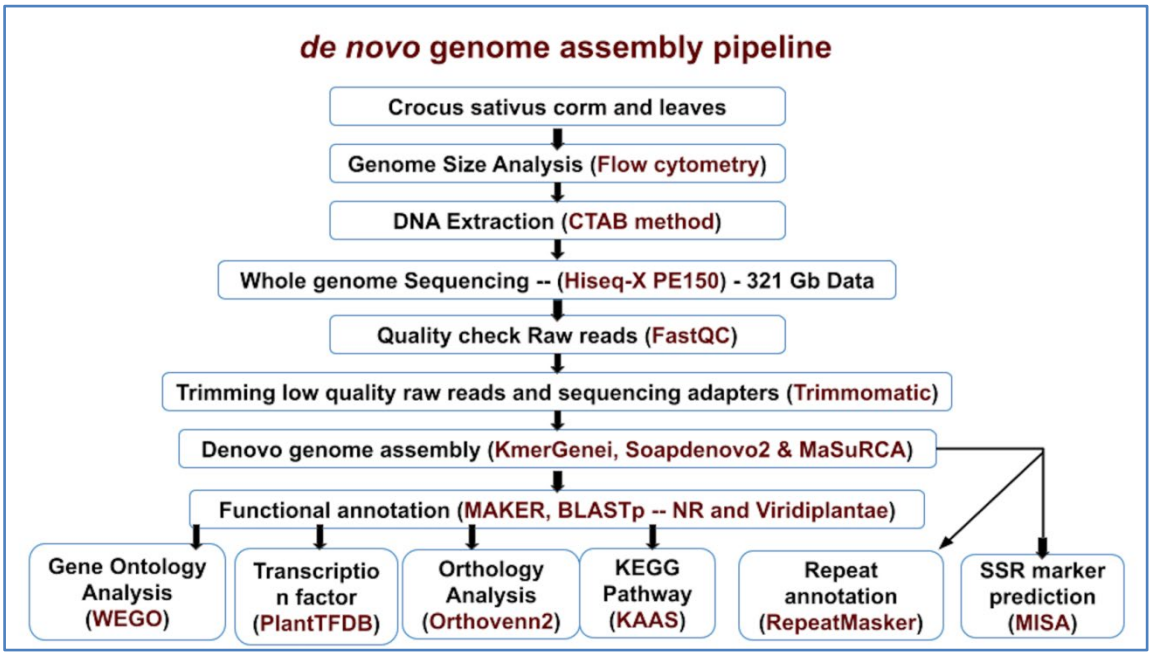


Figure 1: Schematic of *de-novo* genome assembly and annotation pipeline. Black colour text represents the analytical processes and Red colour text represents the software/instrument used to perform the processes.

Materials and Methods:
C. sativus corms were collected from Kishtwar, J&K (33.3116° N, 75.7662° E) in 2019. Corms were grown in the pots for period of three months and leaves were harvested for genome size estimation. Genome size of the plant was estimated by flow cytometric (Hare and Johnston 2011) and k-mer based method using Jelly Fish [18]. Genomic DNA was extracted from corm tissue using CTAB method [19] and quality and quantity was accessed using Qubit (Invitrogen) and agarose gel electrophoresis. 3 microgram DNA was used to construct WGS DNA libraries with 550bp and 800bp insert sizes using NEB next Ultra DNA Library Preparation Kit according to the Illumina's protocol. Quality of the libraries was evaluated using Tapestation (Agilent 4200) and Qubit HS DNA Assay Kit (Invitrogen) and sequenced on HiSeqX platform (150-bp paired-end (PE) reads) to generate 321 Gb data (~92X coverage). Quality of raw reads was evaluated using FastQC tool [20] and low quality bases (<q30) and sequencing adapters were removed using trimmomatic software [21]. *De-novo* genome assembly was performed using Soapdenovo2 [22] and MaSuRCA

[23]. Soapdenovo2 assembly was executed using different kmers (73 kmer predicted by KmerGenei along with 69, 71 kmers) [24]. The statistics of soapdenovo2 assemblies were compared to select the better assembly that was designated as Cs_Assembly_1. MaSuRCA assembly was done using the raw reads and was designated as Cs_Assembly_2. The quality of assemblies was accessed using BUSCO against Viridiplantae lineage from OrthoDB database [25]. Subsequently, raw illumina reads were mapped back to Cs_Assembly_2 using Bowtie2 [26] and previously published transcriptome data [27, 28] was mapped to Cs_Assembly_2 using BWA [29].

Repetitive regions in Cs_Assembly_2 was identified using Repeatmasker and GenomeScope v2 [30, 31] and SSR markers were identified using MISA [32]. Cs_Assembly_2 was further analysed for gene prediction using the MAKER [33] wherein *C. sativus* transcriptome data was used as EST evidence [28], Viridiplantae database (UNIPROT) as protein evidence, maize as Augustus gene prediction model and *Oryza sativa* as snap hmm. Predicted proteins were further annotated using BLASTp against NR (NCBI) and viridiplanteae (UNIPROT) database with modified parameters (E-value-1e^{-3}, sequence identity >40% and query coverage >70%). Annotated proteins were analysed for GO annotations against biological processes, cellular component and metabolic processes using WEGO [34]. Transcription factors (TFs) proteins were identified against PlantTFDB [35] using BLASTp with the modified parameters (E-value-1e^{-3}, sequence identity >30%, query coverage >70%). Orthologous genes were comparedwith *Asparagus officinalis*, *Phalaenopsis equestris*, *Apostatia shenzhenica* of the same plant order along with *Oryza sativa* (Rice) using Orthovenn2 [36]. The proteins sequences of all the plants were downloaded from Phytozome database [37]. Various metabolic pathways *in C. sativus genome* were analysed using KAAS webserver [38].

Data availability:

Whole genome sequencing raw reads and draft genome of *Crocus sativus* has been submitted to NCBI SRA under bioproject PRJNA734464 and PRJNA739096 respectively. All the processed data including draft genome, annotated proteins, and supplementary tables can be accessed at CAPS_NCBS server [39].

Table 1: Assembly statistics of *C. sativus* genome using soapdenovo2 and MaSuRCA *de-novo* assemblers

	Soapdenovo2		MaSuRCA	
Assemblies	-	**Cs_Assembly_1**	-	**Cs_Assembly_2**
kmers	69	71	73	**99**
N50 Scaffold (bases)	1443	1596	1508	**1863**
Number of Scaffolds	1537310	1505129	1433675	**2564042**
Largest Scaffold (bases)	45973	45973	43370	**46734**
Total sequence length	2684437407	2787926280	2589039086	**3014612563**
GC%	43.2	43.2	43.2	**43.2**
Genome Coverage (%)	75.01%	77.90%	72.34%	**84.24%**
BUSCO (%)	7.32%	7.81%	7.05%	**44.46%**

Results & Discussion:

Crocus sativus genome is the first draft genome sequence of the plant belonging to the *Iridaceae* family. Genome size of *C. sativus* was estimated to be 3.5 Gb (3,578,575,507 bases), using flow cytometry and kmer method. Genome size estimated was comparable to earlier reports, wherein it was estimated to be 3.44 Gb using flow cytometry being grown in Italy, Spain and Israel [8, 9]. On the basis of size of the genome, 321 Gb WGS data of *C. sativus* was generated, with an overall coverage of ~92X using Illumina sequencing (Supplementary Table 1). *De-novo* genome assembly and annotation of *C. sativus* was performed using the bioinformatics pipeline represented in Figure 1. *De-novo* genome assembly using Soapdenovo2 with kmer 71 was comparatively better than other two kmers (69 and 73) and was designated as Cs_Assembly_1 with N50 value of 1596 and 77.9% genome coverage (Table 1). De-novo genome assembly with MaSuRCA was designated as Cs_Assembly_2 with N50 value of 1860 and 84.24% genome coverage. Cs_Assembly_2 was found comparatively better than Cs_Assembly_1 as the assembly statistics, such as N50, largest scaffold, genome coverage and BUSCO completeness were higher in Cs_Assembly_2 than Cs_Assembly_1. (Table 1). Further, ~87.28% of raw reads mapped back to Cs_Assembly_2, thereby indicating that most of data has been utilized for genome assembly. In addition, two previously published transcriptome data sets [28, 29] were mapped to the Cs_Assembly_2 and mapping percentage of 99.92% and 92.02% were observed against Cs_Assembly_2 (Supplementary Table 2). High mapping percentage represented the presence of most of the reported exons/CDS in the Cs_Assembly_2 even though the genome assembly was fragmented with less N50 value. *Polygonum cuspidatum* genome was de-novo assembled using Soapdenovo2 with Illumina reads and generated an assembly of 2.56 Gb, with N50 value of 3215 and 98.5% genome coverage [40]. Similarly, the genome of *Linum usitatissimum,* flax plant was *de-novo* assembled using Illumina reads having N50 scaffold of 694 Kb with 81% of genome coverage [41]. Genome coverage of *C. sativus* was comparatively more than flax genome but less than *Polygonum cuspidatum* genome using same sequencing technologies. Total repeats length in *C. sativus* genome (Cs_Assembly_2) was 1,460,908,750 bp (40.8%) as predicted by Genome Scope version 2. A total of 862,275 repeats were identified in Cs_Assembly_2 wherein simple repeat (48.41%) and LTR (30.34%) were the most abundant in the genome. Specifically, Copia & Gypsy were the most abundant LTR repeats (Supplementary Table 3). A total of 9,64,231 SSR markers were identified in Cs_Assembly_2 wherein monomeric SSR repeats (4,86,140 (50.4%)) were more abundant as compared to dinucleotide (2,94,819 (30.5%)) and trinucleotide repeats (1,46,991 (15.2%)) with "A", "TA","TTG" most abundant SSRs in each groups. The abundance of Tetranucleotide (15,375 (1.59%)), pentanucleotide (8,596 (0.9%)) and hexanucleotides (12,310 (1.27%)) repeats each was less than 2% of

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 1-13 (2022)

total SSRs with "AAAT", "TATAT" and "TAACCC" most abundant in respective SSRs (Supplementary Table 4). SSR markers are reported to be multi-allelic, relatively abundant, widely dispersed across the genome and have been used in genetic diversity analysis, parentage assessment, species identification and mapping genetic linkage [42]. These markers can be further evaluated for their application in *C. sativus*. Earlier studies on *C. sativus* transcriptome have reported the presence of 16,721 SSRs [28] and 79,028 SSRs [43] using transcriptome analysis, but higher number of SSR (964,231) were discovered in the present study based on genome sequence.

In total 254,038 proteins were predicted from Cs_Assembly_2 using MAKER pipeline. A total of 52,435 and 52,545 proteins were annotated based on BLASTp against NR and viridiplanteae database respectively (Supplementary Table 5). BUSCO analysis revealed the presence of 75.7% of the plant conserved genes/orthologues in the *C. sativus* genome. Out of total proteins, 51% (26796) were annotated to 8 top-hit plant species **(**Figure 2*)*. Maximum number of proteins was annotated against *Asparagus officinalis* (9213) indicating *C. sativus* to be phylogenetically closer to *Asparagus officinalis*, as both the plants belong to same plant order Asparagales **(**Figure 2). 85% of total proteins (43,649) were associated with gene ontology (GO) ids and classified into biological processes (BP: 22,092 proteins) abundant in cellular and metabolic processes, cellular components (CC: 24,399 proteins) mostly localised in cell and organelle parts and molecular functions (MF: 34,442 proteins) most abundant in catalytic and transporter activities (Supplementary Table 6). Out of the total annotated proteins, 5726 unique *C. sativus* proteins were identified as transcription factors (TFs) belonging to 57 TFs families. MYB & MYB related family proteins (11.86%), being more abundant TFs, followed by bHLH, C2H2, NAC, FAR1, C3H, ERF, bZIP, WRKY and B3 were the top 10 abundant transcription factors family proteins (Supplementary Figure 1, Supplementary Table 7). TFs like MYB & MYB related, bHLH, WRKY are reported to regulate secondary metabolite (apocarotenoid) biosynthesis in *C. sativus* [28]. Earlier reports on *C. sativus* transcriptome has identified less number of TFs (3819, 2601), whereas the most abundant TFs family remains same [27, 28].

C. sativus annotated proteins (52,545) was compared with 3 monocots plants of same order, whose genome and annotations were available in Phytozome database [37], namely *Asparagus officinalis*, *Phalaenopsis equestris*, *Apostatia shenzhenica* along with a model monocot plant *Oryza sativa* (Rice) using Orthovenn2 (Figure 3). A total of 23,744 proteins cluster were found in all the plants wherein 21,606 were orthologous clusters that were atleast common in two species and 2138 were single copy gene clusters wherein each cluster have only one gene from each plant species. Conservation of 7328 proteins clusters, comprising of 51,803 proteins, was observed among the five species (*C. sativus*: 10,001 proteins, *A. officinalis*: 9552, *P. equestris*: 9012, *A. shenzhenica*: 8570 and *O. sativa*: 14,668) (Supplementary Figure 2). The conserved proteins clusters were found to be associated with biological processes (BP-23,010 proteins), cellular component (CC-582 proteins) and molecular functions (MF-957 proteins) and were enriched in defence response, RNA modification, DNA integration, regulation of transcription, rRNA processing and protein phosphorylation (Supplementary Table 8). However, 2510 protein clusters (7914 proteins) were unique to *Crocus sativus* only, out of which 1636 clusters (4595 proteins) were associated with slimmed GO terms (BP: 5201, CC: 63, MF:303 proteins) associated with nucleic acid binding, transferase, hydrolase, oxidoreductase activity and protein and DNA binding activity (Supplemenatary Table 8). As per orthology analysis also, *C. sativus* was found phylogenetically closer to *A. officinalis* as more protein clusters were orthologous between *Crocus sativus* and *Asparagus officinalis* than to other plants compared in the study (Supplementary Figure 3).

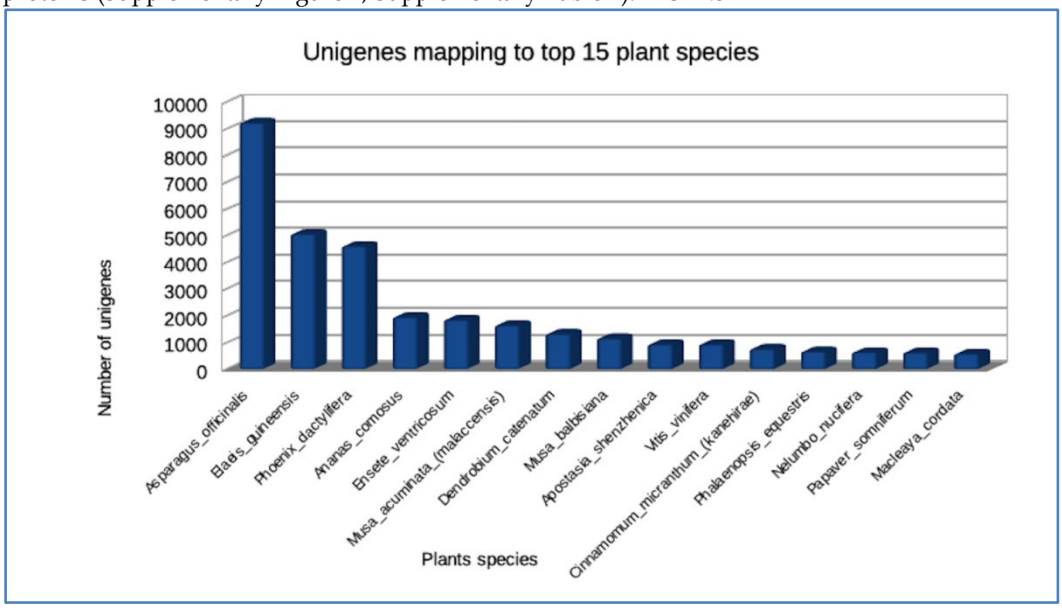


Figure 2: *Crocus sativus* unigenes mapping to top 15 plant species wherein most of the proteins annotated against *Asparagus officinalis*.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 1-13 (2022)

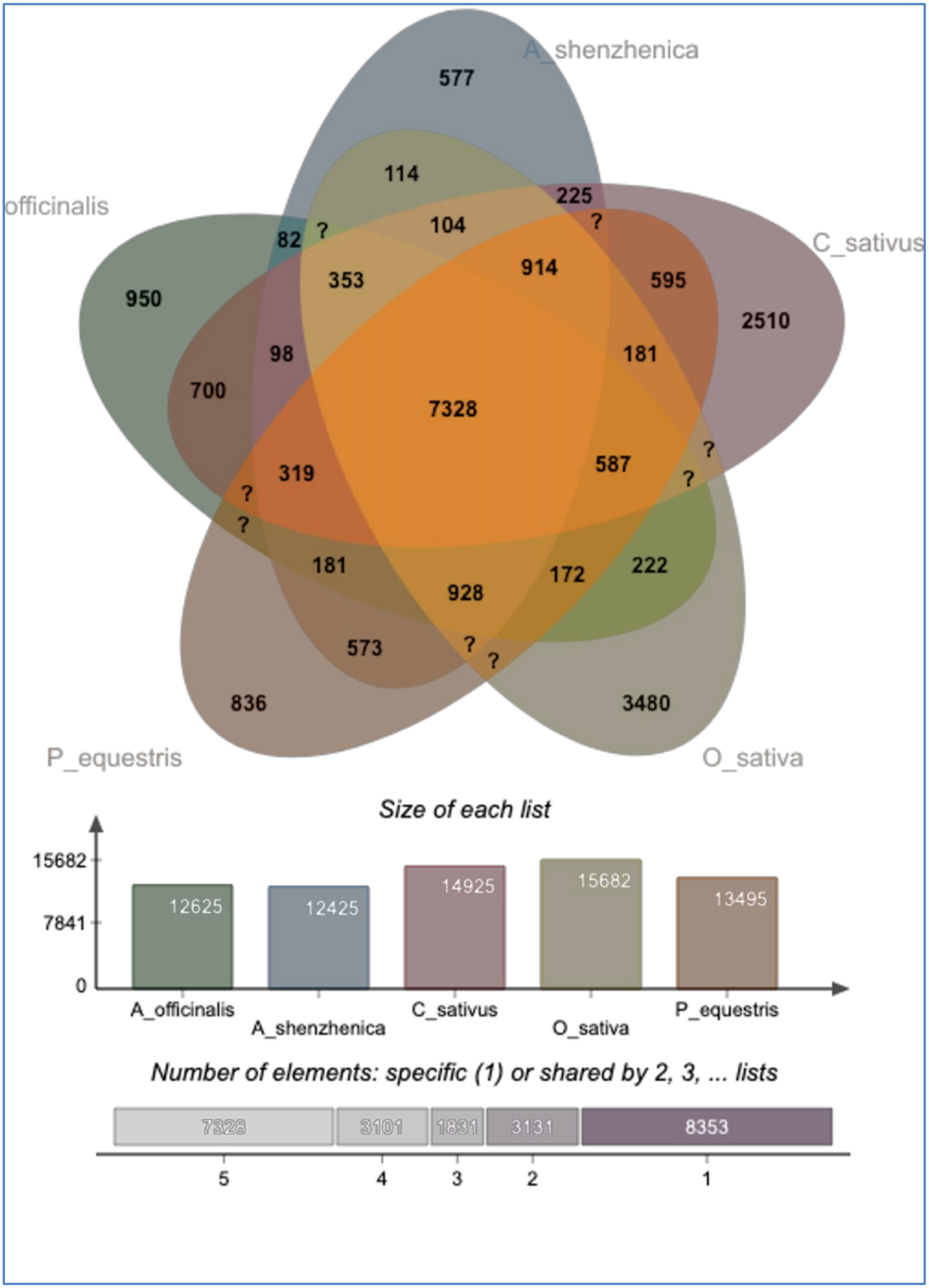


Figure 3: Orthology analysis of *Crocus sativus* with neighboring plants from same order (*Asparagus officinalis*, *Phalaenopsis equestris*, *Apostatia shenzhenica)* along with *Oryza sativa* representing 7328 proteins clusters to be conserved in all the five plant species, whereas 2510 proteins cluster were unique to *C. sativus only*.

A total of 10,912 *C. sativus* proteins were mapped to 395 KEGG pathways of monocots. Various pathways like carbohydrate metabolism, energy metabolism, lipid metabolism, nucleotide metabolism, amino acid metabolism, glycan metabolism, metabolism of cofactors and vitamins along with biosynthesis of terpenoids, polyketides and other secondary metabolites were found complete wherein all the genes involved in pathway were present in draft assembly. We further investigated the presence of genes involved in the synthesis of apocarotenoids namely crocins, picrocrocin, and safranal that are produced in the stigma of *C. sativus*. These apocarotenoids impart red color, bitter taste, and pungent aroma to stigma of saffron and have various medicinal properties [7]. The molecular basis of apocarotenoid biosynthesis in *C. sativus* has been well studied using transcriptomics studies [27, 28]. In the present study, the genes encoding the enzymes involved in carotene biosynthesis pathway, regulating the apocarotenoids synthesis, were present in the *C. sativus* genome (Supplementary Figure 4). This is the first *de-novo* draft genome sequence of *Crocus sativus* that needs to be complemented with the long read sequencing technology (PacBio) to fill in the gaps in the present genome to generate a complete genome sequence. However, this draft genome sequence, in addition to revealing previous unknown genomic information on saffron, will be used as a reference genome for future genome sequencing attempts in saffron.

Conclusion:

It is of interest to establish a *de-novo* reference genome of *Crocus sativus* for the first time. *De-novo* assembly of *Crocus sativus* has been constructed using only Illumina short read, thus, has large number of scaffolds and assembly gaps thereby indicating that our assembly should be referred to as a draft genome sequence. Nevertheless, this study represents the first attempt to assemble the *Crocus sativus* genome, providing a valuable resource for the community to facilitate future research.

Conflict of Interest statement:

The authors declare no conflict of Interest.

Acknowledgments:

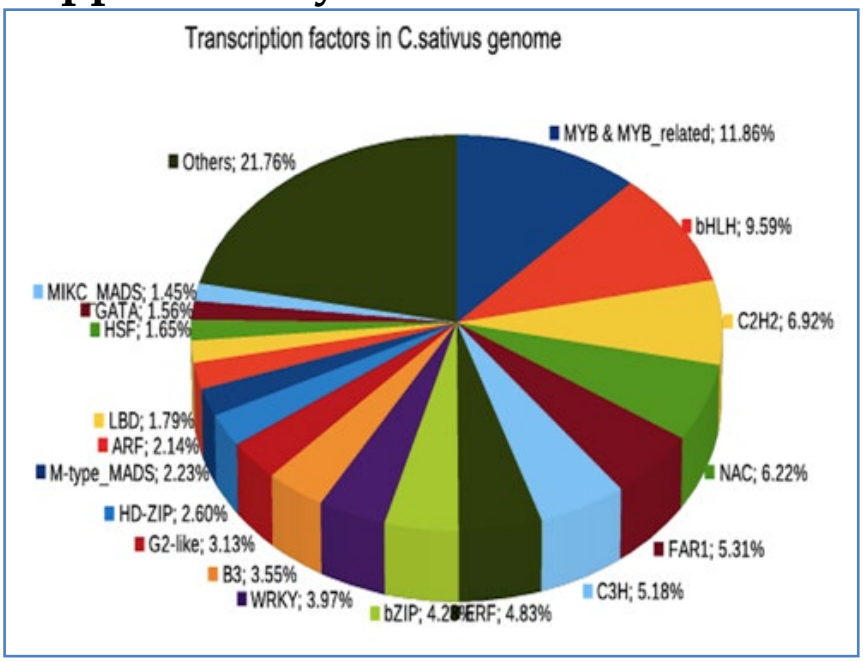
Authors are grateful to Mr. Shanu Magotra, School of Biotechnology, University of Jammu, J &K for his help in sample collection from Kishtwar, Jammu and Kashmir. Authors are thankful to NCBS CIFF facility for their help in flow cytometry analysis. SA is thankful to DST-Women Scientist-A research grant (SR/WOS-A/LS-96-2018) for funding this research. RS acknowledges NCBS (TIFR) and funding through her Mazundar-Shaw Chair in Computational Biology at Institute of Bioinformatics and Applied Biotechnology, Bengaluru, India.

References:

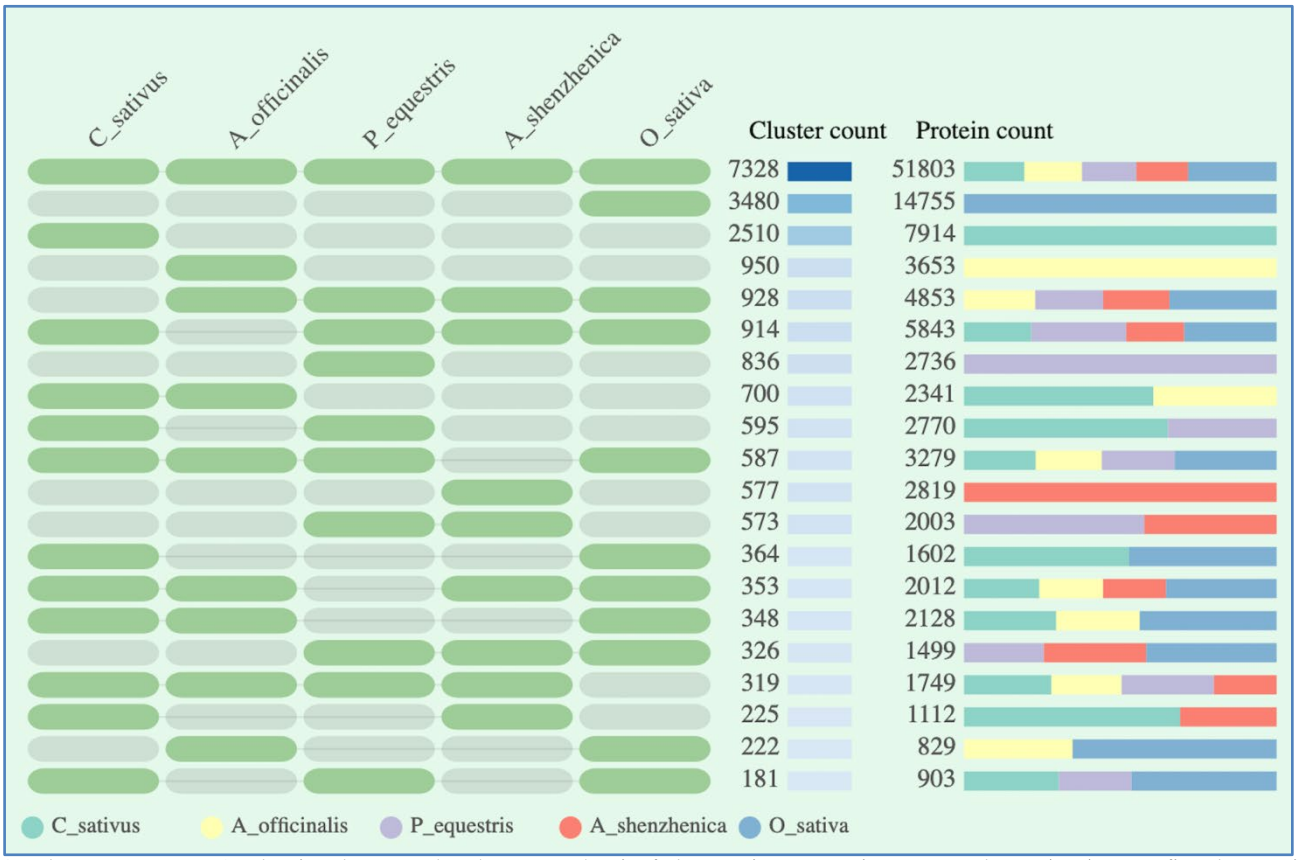
[1] Choi JY *et al. Genome Biol.* 2020, **21**: 21 [PMID: **32019604]**
[2] Liu J *et al. Genome Biol.* 2020**21**: 121 https://doi.org/10.1186/s13059-020-02029-9
[3] Harkess A *et al. Nat Commun.* 2017**8**: 1279 [PMID: **29093472]**
[4] Alonge M *et al. Genetics*. 2020 **216**: 599 [PMID: 32796007]
[5] Xia E *et al. Mol Plant*.2020 **13**: 1013 [PMID: **32353625]**
[6] Magotra S *et al. Sci Rep*.2021 **11**: 5454. [PMID: **33750799]**
[7] Maggi MA *et al. Molecules* 2020 **25**: 5618 [PMID: **33260389]**
[8] Brandizzi F and Caiola MG, *Giornale Bot. Italiano*.1996 **130**: 643
[9] Brandizzi F and Caiola MG, *Plant Syst. Evol*. 1998 **211**: 149
[10] Rubio-Moraga A *et al. BMC Res. Notes*. 2009 **2**: 189 [PMID: **19772674]**
[11] Siracusa L *et al. Genet. Resour. Crop Evol*. 2013 **60**: 711. DOI: 10.1007/s10722-012-9868-9
[12] Busconi M *et al. Plant Sci*.2018 **277**: 1 [PMID: **30466573]**
[13] Mir MA *et al. Saudi Journal of Biological Sciences* 2021 **28**: 1308 [PMID: **33613060]**
[14] Alsayied NF *et al. Ann Bot.* 2015 **116**: 359 [PMID: 26138822]
[15] Nemati Z *et al. Mol Phylogenet Evol*. 2019 **136:** 14 [PMID: **30946897]**
[16] Edwards D and Batley J,*Plant Biotech.*2010 **8**: 2 [PMID: **19906089]**
[17] https://www.biorxiv.org/content/10.1101/2021.06.23.449592v1
[18] Marçais G and Kingsford C,*Bioinformatics*. 2011**27**: 764 [PMID: **21217122]**
[19] Rogers SO and Bendich AJ. In *Plant molecular biology manual* (ed. S. B. Gelvin and R. A. Schilperoort) 1994. 2nd edition, vol. D1:1–8. https://doi.org/10.1007/978-94-011-0511-8_12
[20] Andrews S. 2010 http://www.bioinformatics.babraham.ac.uk/projects/fastqc
[21] Bolger AM *et al. Bioinforma Oxf Engl.*2014 **30:** 2114[PMID: **24695404]**
[22] Luo R *et al. Gigascience.*2012 **1**: 18 [PMID: **23587118]**
[23] Zimin AV *et al. Bioinformatics.* 2013 **29**: 2669 [PMID: **23990416]**
[24] Chikhi R and Madvedev P, *Bioinformatics.* 2013 **30:** 31 [PMID: **23732276]**
[25] Simao FA *et al. Bioinformatics.* 2015 **31:** 3210 [PMID: **26059717]**
[26] Langmead B Curr. Protoc. *Bioinforma.* 2010 Chapter11**:** 11.7.1 [PMID: **21154709]**
[27] Baba SA *et al.* BMC *Genomics.*2015 **16**: 698 [PMID: **26370545]**
[28] Jain M *et al.*, *Sci Rep.*2016 **6**: 22456. [PMID: **26936416]**
[29] Li H and Durbin R, *Bioinformatics.*2010 **26**: 589 [PMID: **20080505]**
[30] Ranallo-Benavidez TR *et al. Nat Commun.* 2020 **11**:1432 [PMID: **32188846]**
[31] Smit AFA *et al.*, 2010 http://www.repeatmasker.org
[32] Beier S *et al. Bioinformatics*. 2017 **33**: 2583 [PMID: **28398459]**
[33] Campbell MS *et al. Curr Protoc. Bioinforma*. 2014 **48**:4.11.1 [PMID: **25501943]**
[34] Jia Y *et al. Nucleic Acids Research.* 2018 **46**: W71 [PMID: **29788377]**
[35] Jinpu J *et al. Nucleic Acids Research.* 2017 **45**: D1040 [PMID: **27924042]**
[36] Xu L *et al. NucleicAcidsRes*. 2019**47**: W52 [PMID: 31053848]
[37] Goodstein DM *et al. Nucleic Acids Res.*2012 **40**: D1178 [PMID: **22110026]**
[38] Moriya Y *et al. NucleicAcidsRes.* 2007 **35**:W182 [PMID: 17526522]
[39] http://caps.ncbs.res.in/download/csat.
[40] Zhang Y *et al. FrontPlantSci*. 2019 **10**:1274 [PMID: **31681373]**
[41] Wang Z *et al. Plant J*.2012 **72**: 461 [PMID: **22757964]**
[42] Feng S *et al. Front. Genet.* 2016 **7**:113 [PMID: **27379163]**
[43] Qian X *et al. BMC genomics.* 2019 **20:** 857 [PMID: **31726972]**

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 1-13 (2022)

Supplementary materials

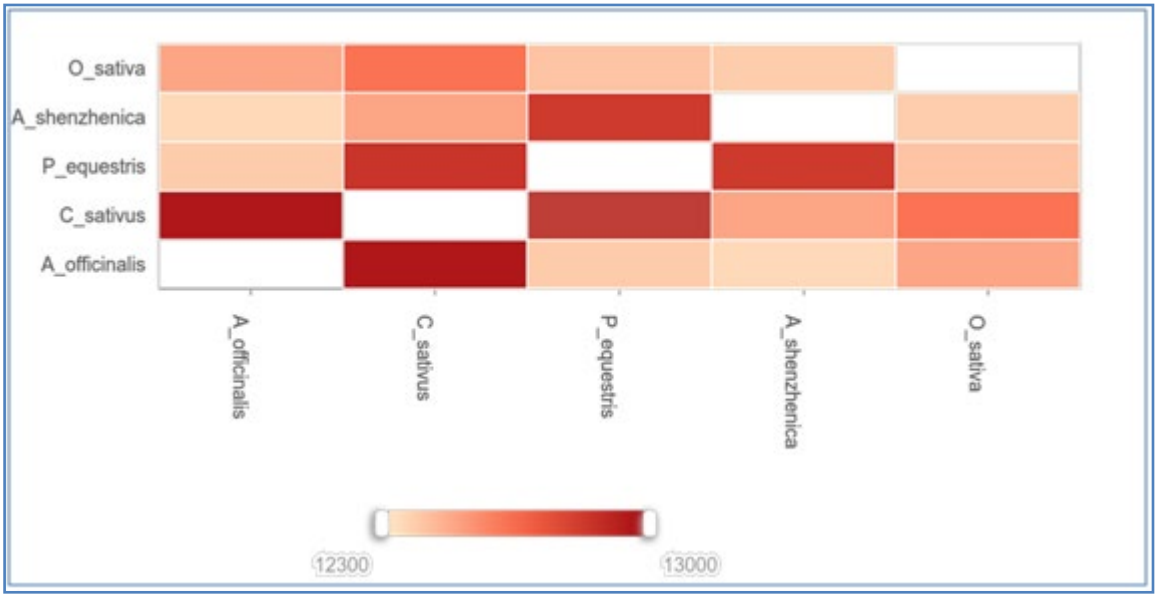


Supplementary Figure 1: Trancription factors identified in *Crocus sativus* genome wherein MYB & MYB related TFs were most abundant.

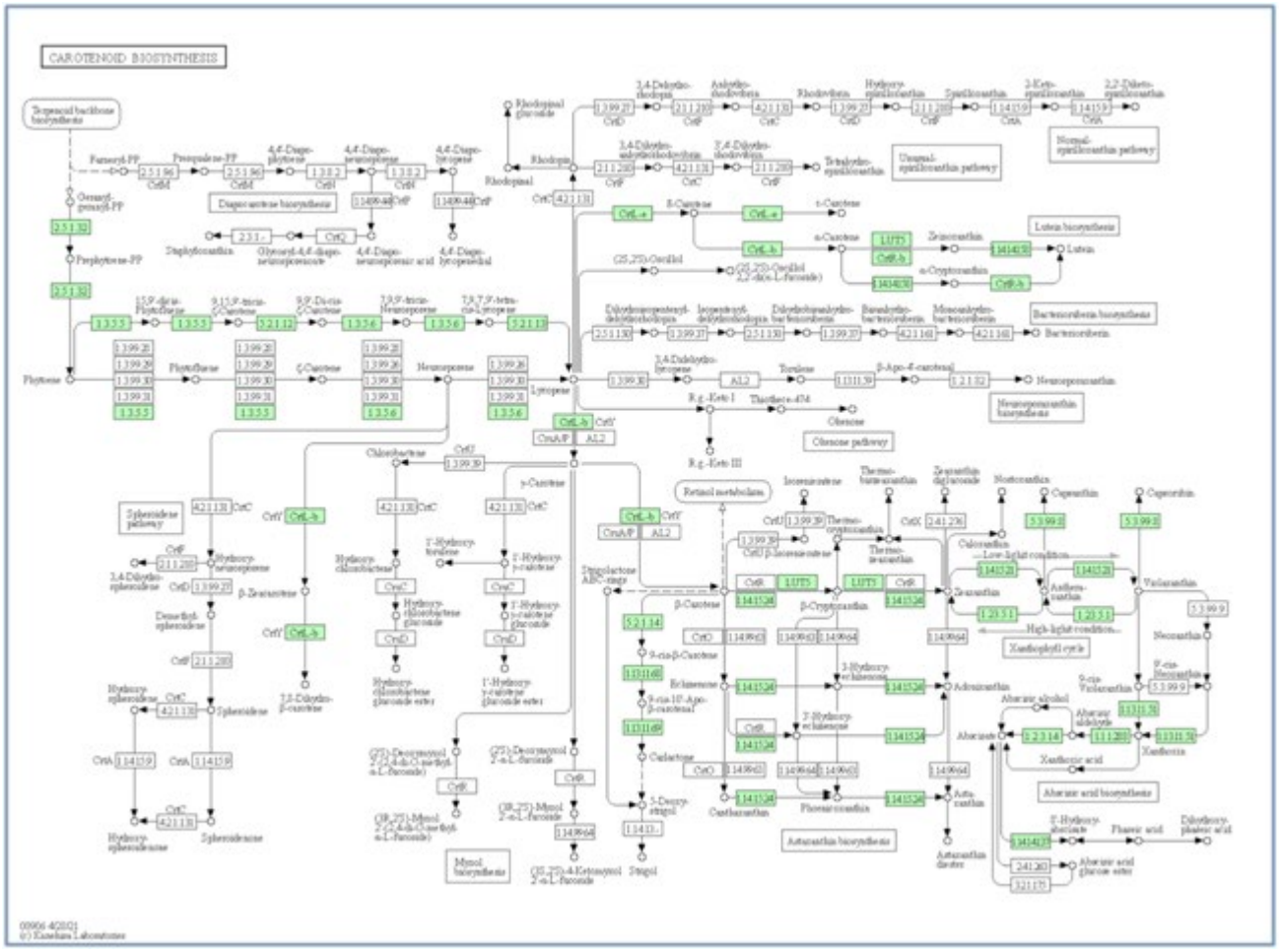


Supplementary Figure 2: **Overlapping cluster numbers between each pair of plant species representing common clusters (7328) among five plant species and unique cluster (2510) to *Crocus sativus*.**

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 1-13 (2022)



Supplementary Figure 3: Heatmap **of overlapping cluster numbers between each pair of plant species representing more number of overlapping clusters between *Crocus sativus* and *Asparagus officinalis*.**



Supplementary Figure 4: Carotene biosynthesis pathway in *Crocus sativus* that was found complete with all the genes in the pathway.

ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 18(1): 1-13 (2022)

Supplementary Table 1: Total raw data of 321.36 Gigabases obtained from two insert size (500 bp and 800 bp) libraries using Illumina sequencing with an overall coverage of ~92X. The genome size was estimated as 3.5 Gigabases (3,578,575,507 bases).

Libraries insert Size	Sample Name	% GC	Sequences (Million Read)	**Data (in Gigabases)**
800	Lib1_800_R1	46%	93.2	13.98
	Lib1_800_R2	46%	93.2	13.98
	Lib2_800_R1	46%	357.8	53.68
	Lib2_800_R2	46%	357.8	53.68
500	Lib3_500_R1	46%	300.1	45.02
	Lib3_500_R2	46%	300.1	45.02
	Lib4_500_R1	46%	320.0	48.01
	Lib4_500_R2	46%	320.0	48.01
Total			**2142.2**	**321.36**

Supplementary Table 3: Classification of repetitive sequences in *C. sativus* genome representing abundance of Simple repeats and LTRs.

Repetitive region	**Numbers**
Simple repeats	415561
LTR	260472
Low_complexity	64624
DNA	64205
LINE	45739
Satellite	3946
RC_Helitron	3072
rRNA	2749
SINE	848
tRNA	650
Other	340
snRNA	54
Retroposon	15
Total	**862275**

Supplementary Table 7**: Transcription factors identified in *C.sativus* genome representing more relative abundance of MYB & MYB related TFs as compared to others.**

Transcription factors	**Numbers**	**Percentage(%)**
MYB & MYB_related	648	11.32
bHLH	524	9.15
C2H2	378	6.6
NAC	340	5.94
FAR1	290	5.06
C3H	283	4.94
ERF	264	4.61
bZIP	234	4.09
WRKY	217	3.79
B3	194	3.39
G2-like	171	2.99
HD-ZIP	142	2.48
GRAS	133	2.32
Trihelix	128	2.24
M-type_MADS	122	2.13
ARF	117	2.04
LBD	98	1.71
HSF	90	1.57
GATA	85	1.48
MIKC_MADS	79	1.38
HB-other	77	1.34
TALE	65	1.14
Nin-like	65	1.14
Dof	61	1.07
TCP	58	1.01
AP2	58	1.01
NF-YC	55	0.96
CAMTA	51	0.89
SBP	49	0.86
BES1	49	0.86
NF-YB	47	0.82
ARR-B	41	0.72
ZF-HD	40	0.7
E2F/DP	38	0.66
GeBP	37	0.65
NF-YA	35	0.61
CPP	34	0.59
CO-like	33	0.58
WOX	30	0.52
GRF	29	0.51
DBB	28	0.49
YABBY	23	0.4
S1Fa-like	20	0.35
EIL	20	0.35
BBR-BPC	20	0.35
SRS	19	0.33
LSD	18	0.31
RAV	15	0.26
NF-X1	15	0.26
STAT	14	0.24
Whirly	13	0.23
VOZ	9	0.16
HRT-like	8	0.14
HB-PHD	8	0.14
LFY	5	0.09
SAP	2	0.03
Total	**5726**	**100**

Supplementary Table 2: Mapping WGS raw reads and previous published transcriptome data to Cs_Assembly_2.

Data Type	**WGS raw reads in present study**	***C.sativus* transcripts (Jain et al., 2016)**	***C.sativus* transcriptome raw reads (Baba et al 2015)**
Source of Data	**Present study**	**Jain et al., 2016**	**Baba et al 2015**
Total number of reads/transcripts	2135957246	327920	59043670
Mapped reads/transcripts	1864292472	327643	54330850
Mapping Percentage	87.28 %	99.92 %	92.02 %

Supplementary Table 4: SSR markers from *Crocus sativus* draft genome (Cs_Assembly_2) depicting the more relative abundance of monomeric repeat microsatellite.

SSR types	count	relative %age	Most abundant	%age
monomeric repeat microsatellite	486140	50.4%	"A"	44.6 %
dinucleotide repeat microsatellite	294819	30.5%	"TA"	16.5 %
trinucleotide repeat microsatellite	146991	15.2%	"TTG"	5.72 %
Tetranucleotide repeat microsatellite	15375	1.59%	"AAAT"	7.7 %
Pentanucleotide repeat microsatellite	8596	0.9%	"TATAT"	3.2 %
Hexanucleotide repeat microsatellite	12310	1.27%	"TAACCC"	5.3 %
Total	964231			

Supplementary Table 5: Number of genes annotated against NR and Viridiplantae database depicting more number of proteins annotated against Viridiplantae database.

Databases	Total maker annotated	Total annotated proteins	>40% percentage identity &> 70% query coverage	Unique accession numbers
Viridiplantae	254038	146118	107553	52546
NR	254038	143745	107385	52436

Supplementary Table 6: **Total number of annotated genes associated with total GO terms, Biological Process, Cellular Components and Molecular Functions.**

Annotated Genes associated with Gene Ontology (GO)	**Total number**	**GO terms Biological Process**				**GO terms Cellular Component**			
GO Terms	43649	**GO Ids**	**Gene number**	**Percentage**	**Term description**	**GO Ids**	**Gene number**	**Percentage**	**Term description**
Biological Processes	22092	GO:0009987	17240	39.5	cellular process	GO:0005623	15718	36.0	cell
Cellular component	24399	GO:0008152	16215	37.1	metabolic process	GO:0044464	15540	35.6	cell part
Molecular Function	34442	GO:0065007	4111	9.4	biological regulation	GO:0044422	4914	11.3	organelle part
		GO:0050789	3474	8	regulation of biological process	GO:0043226	12294	28.2	organelle
GO terms Molecular Function		GO:0051179	3028	6.9	localization	GO:0044425	12152	27.8	membrane part
GO Ids	**Gene number**	**Percentage**	**Term description**	GO:0071840	2814	6.4	cellular component organization or biogenesis	GO:0016020	13123	30.1	membrane
GO:0003824	21143	48.4	catalytic activity	GO:0050896	2592	5.9	response to stimulus	GO:0032991	5405	12.4	protein-containing complex
GO:0005215	2226	5.1	transporter activity	GO:0023052	936	2.1	signaling	GO:0099080	313	0.7	supramolecular complex
GO:0005488	21987	50.4	binding	GO:0032502	605	1.4	developmental process	GO:0031974	920	2.1	membrane-enclosed lumen
GO:0140104	62	0.1	molecular carrier activity	GO:0048519	507	1.2	negative regulation of biological process	GO:0005576	559	1.3	extracellular region
GO:0060089	186	0.4	molecular transducer activity	GO:0032501	496	1.1	multicellular organismal process	GO:0044421	66	0.2	extracellular region part
GO:0098772	700	1.6	molecular function	GO:0048518	489	1.1	positive regulation of	GO:0009295	16	0.0	nucleoid

			regulator				biological process				
GO:0045735	41	0.1	nutrient reservoir activity	GO:0000003	317	0.7	reproduction	GO:0055044	107	0.2	symplast
GO:0016209	224	0.5	antioxidant activity	GO:0022414	316	0.7	reproductive process	GO:0030054	110	0.3	cell junction
GO:0140110	883	2.0	transcription regulator activity	GO:0051704	186	0.4	multi-organism process	GO:0044217	141	0.3	other organism part
GO:0038024	7	0.0	cargo receptor activity	GO:0019740	117	0.3	nitrogen utilization	GO:0044215	141	0.3	other organism
GO:0045182	5	0.0	translation regulator activity	GO:0040007	69	0.2	growth	GO:0019012	17	0.0	virion
GO:0005198	1315	3.0	structural molecule activity	GO:0002376	57	0.1	immune system process	GO:0044423	17	0.0	virion part
GO:0031386	6	0.0	protein tag	GO:0098754	33	0.1	detoxification				
				GO:0048511	22	0.1	rhythmic process				
				GO:0040011	13	0	locomotion				
				GO:0008283	10	0	cell proliferation				
				GO:0015976	9	0	carbon utilization				
				GO:0022610	9	0	biological adhesion				
				GO:0043473	3	0	pigmentation				

Supplementary Table 8: **Orthology analysis of *C. sativus* with 3 monocots plants of same order namely *Asparagus officinalis, Phalaenopsis equestris, Apostatia shenzhenica* along with a model monocot plant *Oryza sativus* (Rice) using Orthovenn2. Common core protein clusters and unique proteins clusters associated with Biological processes, cellular component, and Molecular functions have been identified.**

Total Proteins and clusters				Common_cluster_GO_enrichment				
Species	**Proteins**	**Clusters**	**Singletons**	**GO_IDs**	**Processes**	**Protein count**	**GO_Terms**	**P-value**
C_sativus	52988	14925	24611	GO:0006952	defense response	26	biological_process	1.79E-11
A_officinalis	27395	12625	8637	GO:0009451	RNA modification	73	biological_process	1.44E-09
P_equestris	29415	13495	9391	GO:0015074	DNA integration	2	biological_process	1.41E-08
A_shenzhenica	21743	12425	4275	GO:0006355	regulation of transcription, DNA-templated	124	biological_process	4.57E-06
O_sativa	52424	15682	13175	GO:0006364	rRNA processing	56	biological_process	1.19E-05

Common_cluster_bio_process					
GO_IDs	**Processes**	**Protein count**	**GO_IDs**	**Processes**	**Protein count**
GO:0000003	reproduction	199	GO:0015833	peptide transport	2
GO:0000280	nuclear division	53	GO:0015849	organic acid transport	10
GO:0000746	conjugation	1	GO:0015893	drug transport	5
GO:0001775	cell activation	2	GO:0015931	nucleobase-containing compound transport	22
GO:0002376	immune system process	43	GO:0015976	carbon utilization	2
GO:0005975	carbohydrate metabolic process	214	GO:0015979	photosynthesis	59
GO:0005976	polysaccharide metabolic process	67	GO:0016032	viral process	29
GO:0006066	alcohol metabolic process	7	GO:0016043	cellular component organization	315
GO:0006081	cellular aldehyde metabolic process	14	GO:0016049	cell growth	43
GO:0006082	organic acid metabolic process	262	GO:0016050	vesicle organization	21
GO:0006091	generation of precursor metabolites and energy	52	GO:0016070	RNA metabolic process	746
GO:0006112	energy reserve metabolic process	2	GO:0016192	vesicle-mediated transport	136
GO:0006119	oxidative phosphorylation	8	GO:0016458	gene silencing	33
GO:0006139	nucleobase-containing compound metabolic process	503	GO:0017144	drug metabolic process	2
GO:0006259	DNA metabolic process	90	GO:0019538	protein metabolic process	273
GO:0006260	DNA replication	37	GO:0019748	secondary metabolic process	26

Common_cluster_Cellular_component		
GO_IDs	**Processes**	**Protein count**
GO:0005575	cellular_component	97
GO:0005576	extracellular region	7
GO:0005622	intracellular	43
GO:0005634	nucleus	18
GO:0005730	nucleolus	4
GO:0005739	mitochondrion	67
GO:0005773	vacuole	9
GO:0005783	endoplasmic reticulum	5
GO:0005794	Golgi apparatus	4
GO:0005840	ribosome	3
GO:0005856	cytoskeleton	3
GO:0009536	plastid	23
GO:0009579	thylakoid	11
GO:0016020	membrane	122
GO:0030313	cell envelope	1
GO:0030684	preribosome	1

Unique_cluster_bio_process					
GO_IDs	**Processes**	**Protein count**	**GO_IDs**	**Processes**	**Protein count**
GO:0000003	reproduction	52	GO:0015893	drug transport	1
GO:0000280	nuclear division	3	GO:0015931	nucleobase-containing compound transport	7
GO:0000746	conjugation	1	GO:0015976	carbon utilization	2
GO:0001816	cytokine production	1	GO:0015979	photosynthesis	11
GO:0002376	immune system process	7	GO:0016032	viral process	7
GO:0005975	carbohydrate metabolic process	66	GO:0016043	cellular component organization	71
GO:0005976	polysaccharide metabolic process	34	GO:0016049	cell growth	20
GO:0006066	alcohol metabolic process	5	GO:0016050	vesicle organization	10
GO:0006081	cellular aldehyde metabolic process	2	GO:0016070	RNA metabolic process	125
GO:0006082	organic acid metabolic process	75	GO:0016192	vesicle-mediated transport	42
GO:0006091	generation of precursor metabolites and energy	21	GO:0016458	gene silencing	1
GO:0006119	oxidative phosphorylation	9	GO:0017144	drug metabolic process	2
GO:0006139	nucleobase-containing compound metabolic process	94	GO:0019538	protein metabolic process	59
GO:0006259	DNA metabolic process	14	GO:0019748	secondary metabolic process	11
GO:0006260	DNA replication	3	GO:0022406	membrane docking	3
GO:0006281	DNA repair	5	GO:0022411	cellular component disassembly	6

GO:0006281	DNA repair	46	GO:0022406	membrane docking	10		GO:0031982	vesicle	1		GO:0006304	DNA modification	1	GO:0022607	cellular component assembly	9
GO:0006304	DNA modification	12	GO:0022411	cellular component disassembly	20		GO:0032991	macromolecular complex	1		GO:0006323	DNA packaging	1	GO:0031023	microtubule organizing center organization	1
GO:0006323	DNA packaging	8	GO:0022607	cellular component assembly	53		GO:0042579	microbody	1		GO:0006396	RNA processing	47	GO:0031640	killing of cells of other organism	2
GO:0006354	DNA-templated transcription, elongation	7	GO:0031023	microtubule organizing center organization	2		GO:0042597	periplasmic space	1		GO:0006412	translation	46	GO:0032196	transposition	3
GO:0006396	RNA processing	341	GO:0031647	regulation of protein stability	1		GO:0043226	organelle	26		GO:0006457	protein folding	6	GO:0032501	multicellular organismal process	102
GO:0006412	translation	208	GO:0032501	multicellular organismal process	372		GO:0043229	intracellular organelle	26		GO:0006464	cellular protein modification process	91	GO:0032502	developmental process	118
GO:0006457	protein folding	31	GO:0032502	developmental process	449		GO:0043231	intracellular membrane bounded organelle	1		GO:0006508	proteolysis	50	GO:0032989	cellular component morphogenesis	24
GO:0006464	cellular protein modification process	335	GO:0032989	cellular component morphogenesis	70		GO:0043234	protein complex	4		GO:0006518	peptide metabolic process	12	GO:0034622	cellular macromolecular complex assembly	15
GO:0006508	proteolysis	188	GO:0034622	cellular macromolecular complex assembly	78		GO:0044464	cell part	79		GO:0006629	lipid metabolic process	48	GO:0040007	growth	25
GO:0006518	peptide metabolic process	103	GO:0040007	growth	80		GO:0055044	symplast	6		GO:0006725	cellular aromatic compound metabolic process	116	GO:0042044	fluid transport	2
GO:0006629	lipid metabolic process	188	GO:0040011	locomotion	7						GO:0006766	vitamin metabolic process	10	GO:0042180	cellular ketone metabolic process	2
GO:0006662	glycerol ether metabolic process	9	GO:0042044	fluid transport	2		**Common_cluster_Molecular_function**				GO:0006793	phosphorus metabolic process	83	GO:0042254	ribosome biogenesis	8
GO:0006725	cellular aromatic compound metabolic process	595	GO:0042180	cellular ketone metabolic process	27		**GO_IDs**	**Processes**	**Protein count**		GO:0006807	nitrogen compound metabolic process	161	GO:0042440	pigment metabolic process	4
GO:0006730	one-carbon metabolic process	3	GO:0042254	ribosome biogenesis	115		GO:0000166	nucleotide binding	27		GO:0006810	transport	95	GO:0042445	hormone metabolic process	2
GO:0006766	vitamin metabolic process	36	GO:0042440	pigment metabolic process	33		GO:0001871	pattern binding	2		GO:0006811	ion transport	36	GO:0043062	extracellular structure organization	3
GO:0006793	phosphorus metabolic process	235	GO:0042445	hormone metabolic process	11		GO:0001882	nucleoside binding	14		GO:0006818	hydrogen transport	9	GO:0043094	cellular metabolic compound salvage	5
GO:0006805	xenobiotic metabolic process	2	GO:0042886	amide transport	1		GO:0003674	molecular_function	89		GO:0006865	amino acid transport	4	GO:0043101	purine-containing compound salvage	3
GO:0006807	nitrogen compound metabolic process	808	GO:0043062	extracellular structure organization	9		GO:0003676	nucleic acid binding	134		GO:0006869	lipid transport	4	GO:0043170	macromolecule metabolic process	227
GO:0006810	transport	328	GO:0043094	cellular metabolic compound salvage	20		GO:0003824	catalytic activity	4		GO:0006914	autophagy	3	GO:0043412	macromolecule modification	35
GO:0006811	ion transport	95	GO:0043101	purine-containing compound salvage	2		GO:0004386	helicase activity	2		GO:0006928	movement of cell or subcellular component	4	GO:0043603	cellular amide metabolic process	11
GO:0006818	hydrogen transport	16	GO:0043170	macromolecule metabolic process	1177		GO:0004497	monooxygenase activity	3		GO:0006996	organelle organization	37	GO:0044237	cellular metabolic process	296
GO:0006836	neurotransmitter transport	3	GO:0043412	macromolecule modification	222		GO:0004871	signal transducer activity	1		GO:0007005	mitochondrion organization	5	GO:0044238	primary metabolic process	181
GO:0006865	amino acid transport	12	GO:0043449	cellular alkene metabolic process	1		GO:0005215	transporter activity	76		GO:0007029	endoplasmic reticulum organization	3	GO:0044255	cellular lipid metabolic process	37
GO:0006869	lipid transport	17	GO:0043603	cellular amide metabolic process	94		GO:0005488	binding	55		GO:0007031	peroxisome organization	1	GO:0044419	interspecies interaction between organisms	4
GO:0006914	autophagy	9	GO:0044237	cellular metabolic process	1374		GO:0005496	steroid binding	3		GO:0007033	vacuole organization	3	GO:0045229	external encapsulating structure organization	8
GO:0006928	movement of cell or subcellular component	19	GO:0044238	primary metabolic process	879		GO:0005515	protein binding	36		GO:0007049	cell cycle	14	GO:0045333	cellular respiration	20
GO:0006949	syncytium formation	1	GO:0044255	cellular lipid metabolic process	156		GO:0008233	peptidase activity	40		GO:0007059	chromosome segregation	4	GO:0046483	heterocycle metabolic process	115

GO:0006996	organelle organization	155	GO:0044419	interspecies interaction between organisms	28	GO:0008289	lipid binding	3	GO:0007154	cell communication	43	GO:0046903	secretion	4
GO:0006997	nucleus organization	9	GO:0045229	external encapsulating structure organization	19	GO:0008641	small protein activating enzyme activity	1	GO:0008037	cell recognition	3	GO:0048284	organelle fusion	8
GO:0007005	mitochondrion organization	40	GO:0045333	cellular respiration	32	GO:0009055	electron carrier activity	4	GO:0008150	biological_process	685	GO:0048285	organelle fission	2
GO:0007029	endoplasmic reticulum organization	6	GO:0046483	heterocycle metabolic process	604	GO:0016209	antioxidant activity	4	GO:0008152	metabolic process	407	GO:0048469	cell maturation	4
GO:0007030	Golgi organization	2	GO:0046490	isopentenyl diphosphate metabolic process	3	GO:0016491	oxidoreductase activity	56	GO:0008283	cell proliferation	2	GO:0048511	rhythmic process	2
GO:0007031	peroxisome organization	7	GO:0046794	transport of virus	3	GO:0016597	amino acid binding	1	GO:0008643	carbohydrate transport	1	GO:0050896	response to stimulus	202
GO:0007033	vacuole organization	11	GO:0046903	secretion	20	GO:0016740	transferase activity	127	GO:0008655	pyrimidine-containing compound salvage	1	GO:0051179	localization	52
GO:0007049	cell cycle	115	GO:0048284	organelle fusion	20	GO:0016787	hydrolase activity	98	GO:0009116	nucleoside metabolic process	18	GO:0051186	cofactor metabolic process	20
GO:0007059	chromosome segregation	41	GO:0048285	organelle fission	10	GO:0016829	lyase activity	4	GO:0009117	nucleotide metabolic process	23	GO:0051234	establishment of localization	98
GO:0007154	cell communication	145	GO:0048469	cell maturation	19	GO:0016853	isomerase activity	13	GO:0009225	nucleotide-sugar metabolic process	2	GO:0051258	protein polymerization	1
GO:0007155	cell adhesion	3	GO:0048511	rhythmic process	14	GO:0016874	ligase activity	1	GO:0009308	amine metabolic process	3	GO:0051261	protein depolymerization	5
GO:0008037	cell recognition	6	GO:0050877	neurological system process	2	GO:0019213	deacetylase activity	3	GO:0009404	toxin metabolic process	1	GO:0051276	chromosome organization	7
GO:0008150	biological_process	2875	GO:0050896	response to stimulus	786	GO:0019239	deaminase activity	1	GO:0009657	plastid organization	9	GO:0051301	cell division	4
GO:0008152	metabolic process	1906	GO:0051179	localization	197	GO:0019842	vitamin binding	2	GO:0009914	hormone transport	1	GO:0051604	protein maturation	3
GO:0008283	cell proliferation	10	GO:0051181	cofactor transport	5	GO:0030234	enzyme regulator activity	6	GO:0009987	cellular process	337	GO:0051606	detection of stimulus	1
GO:0008643	carbohydrate transport	14	GO:0051186	cofactor metabolic process	115	GO:0030246	carbohydrate binding	6	GO:0010118	stomatal movement	6	GO:0051640	organelle localization	7
GO:0008655	pyrimidine-containing compound salvage	7	GO:0051189	prosthetic group metabolic process	5	GO:0033218	amide binding	1	GO:0015031	protein transport	40	GO:0051641	cellular localization	39
GO:0009116	nucleoside metabolic process	49	GO:0051234	establishment of localization	337	GO:0043021	ribonucleoprotein complex binding	2	GO:0015074	DNA integration	2	GO:0051704	multi-organism process	38
GO:0009117	nucleotide metabolic process	76	GO:0051258	protein polymerization	4	GO:0043167	ion binding	119	GO:0015833	peptide transport	1	GO:0065003	macromolecular complex assembly	9
GO:0009225	nucleotide-sugar metabolic process	16	GO:0051261	protein depolymerization	2	GO:0045735	nutrient reservoir activity	4	GO:0015849	organic acid transport	3	GO:0065007	biological regulation	169
GO:0009292	genetic transfer	2	GO:0051276	chromosome organization	64	GO:0046906	tetrapyrrole binding	1				GO:0071555	cell wall organization	12
GO:0009308	amine metabolic process	25	GO:0051301	cell division	33	GO:0048037	cofactor binding	11						
GO:0009404	toxin metabolic process	7	GO:0051604	protein maturation	33	GO:0051213	dioxygenase activity	2	**Unique_cluster_Molecular_function**					
GO:0009657	plastid organization	34	GO:0051606	detection of stimulus	8	GO:0051540	metal cluster binding	2	**GO_IDs**	**Processes**	**Protein_count**	**GO_IDs**	**Processes**	**Protein_count**
GO:0009914	hormone transport	1	GO:0051640	organelle localization	27	GO:0060089	molecular transducer activity	1	GO:0000166	nucleotide binding	11	GO:0016740	transferase activity	41
GO:0009987	cellular process	1407	GO:0051641	cellular localization	162				GO:0001882	nucleoside binding	10	GO:0016787	hydrolase activity	29
GO:0010118	stomatal movement	17	GO:0051704	multi-organism process	113	**Unique_cluster_Cellular_component**			GO:0003674	molecular_function	32	GO:0016829	lyase activity	1
GO:0010191	mucilage metabolic process	1	GO:0052314	phytoalexin metabolic process	1	**GO_IDs**	**Processes**	**Protein count**	GO:0003676	nucleic acid binding	37	GO:0016853	isomerase activity	3
GO:0015031	protein transport	161	GO:0065003	macromolecular complex assembly	52	GO:0005575	Cellular component	13	GO:0003824	catalytic activity	1	GO:0019842	vitamin binding	1
GO:0015074	DNA integration	2	GO:0065007	biological regulation	768	GO:0005576	extracellular region	5	GO:0004497	monooxygenase activity	3	GO:0030234	enzyme regulator activity	3
GO:0015791	polyol transport	1	GO:0071555	cell wall organization	20	GO:0005622	intracellular	1	GO:0005215	transporter activity	27	GO:0030246	carbohydrate binding	2

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 1-13 (2022)