Data in Brief 31 (2020) 105995

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Transcriptomic dataset reveals the molecular basis of genotypic variation in hexaploid wheat (*T. aestivum* L.) in response to Fe/Zn deficiency

Om Prakash Gupta[a], Vanita Pandey[a], Ritu Saini[a], Sneh Narwal[a], Vipin Kumar Malik[a], Tushar Khandale[a], Sewa Ram[a,*], Gyanendra Pratap Singh[b]

[a] *Division of Quality and Basic Sciences, ICAR-Indian Institute of Wheat and Barley Research (IIWBR), Karnal 132001, Haryana, India*
[b] *Director, ICAR-Indian Institute of Wheat and Barley Research (IIWBR), Karnal 132001, Haryana, India*

ARTICLE INFO

Article history:
Received 23 April 2020
Revised 1 July 2020
Accepted 2 July 2020
Available online 5 July 2020

Keywords:
Fe/Zn deficiency
Hexaploid wheat
Phytosiderophore
Fe/Zn transporters
Biofortification, Sand pot culture

ABSTRACT

The datasets depicted in the paper are related to the original article entitled “Identifying transcripts associated with efficient transport and accumulation of Fe and Zn in hexaploid wheat (*T. aestivum* L.)” [1]. Four wheat genotypes *i.e.* Sonora 64, CRP 1660, Vinata, and DBW 17 were selected for RNA sequencing using Illumina HiSeq4000 platform. These genotypes were grown in Fe/Zn sufficient and deficient conditions in sand pot culture with intermittent administration of Hoagland solution. Pooled assembly was carried out for all of the four varieties subsequent to discarding low-quality reads, adaptor sequences and contamination resulting in approximately 315,904 clean transcripts of around 937 bp lengths and N_{50} of 1,294 bp. For the functional annotation of the identified transcripts databases like Pfam, KEGG pathway, Uniprot, PlnTFDB and wheat proteins were utilized. Differential expression calculation of transcripts was carried out by DESeq, an R package and real-time PCR study of 12 Fe/Zn metabolic pathway related transcripts was utilized for further

DOI of original article: 10.1016/j.jbiotec.2020.03.015
* Corresponding author.
E-mail address: sewa.ram@icar.gov.in (S. Ram).

https://doi.org/10.1016/j.dib.2020.105995

revalidation of data. Elemental analysis of grain Fe and Zn was performed using Flame Atomic Absorption Spectrometry (FAAS). The RNA-seq data of all the four wheat genotypes was uploaded on Sequence Read Archive (SRA: SUB6961770 and BioProject: PRJNA605909), enabling easy access to the researchers worldwide.

Specifications table

Subject	Plant Science
Specific subject area	Wheat Biofortification
Type of data	Tables and Figures
How data were acquired	The data was obtained by Next-generation Sequencing technique utilizing Illumina HiSeq 4000 platform, qPCR, FAAS
Data format	Raw and Analysed
Parameters for data collection	All data were collected from experiments described in methods section conducted in triplicate
Description of data collection	RNA Sequencing was performed using Illumina HiSeq 4000 platform, Genotypic Technology, Bengaluru, India, qPCR was performed using C1000^{TM} Touch Thermal Cycler (CFX96^{TM}, BioRad) FAAS data was collected using FAAS4141 PLUS, ECIL, India
Data source location	ICAR-Indian Institute of Wheat and Barley Research, Karnal-132,001, India
Data accessibility	Raw data of RNA Seq analysis are available on Sequence Read Archive (SRA) database and connectedto SRA: SUB6961770 and BioProject: PRJNA605909 (Direct URL to the data: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA605909) Analyzed data is with the article
Related research article	Gupta OP, Pandey V, Saini R, Narwal S, Malik VK, Khandale T, Ram S, Singh GP, (2020). Identifying transcripts associated with efficient transport and accumulation of Fe and Zn in hexaploid wheat (*T. aestivum* L.). *Journal of Biotechnology,* **316:** 46–55. https://doi.org/10.1016/j.jbiotec.2020.03.015.

Value of the data

- Data presented here would enhance our current knowledge of molecular mechanism of Fe/Zn transportation from rhizosphere up to the grains in wheat.
- Molecular biologists and breeders working in the area of biofortification would get benefit out of this data.
- The data can be used to develop molecular markers for screening wheat genotypes for high and low grain Fe and Zn content.

Data description

The present report contains De-novo transcriptome analysis of four Indian hexaploid wheat varieties (Sonora 64, CRP 1660, Vinata: high, and DBW 17: low) varying in seed Fe/Zn content exposed to Fe/Zn deficiency conditions. Paired-end (PE) reads of 150 bp length were sequenced by an Illumina HiSeq sequencer 4000 which generated a total of 415.03 million (150 × 2) reads, out of which 376.71 million good quality adapter-free reads were used for the downstream analysis. Fig. 1 describes the flowchart for library preparation while Fig. 2A& 2B indicates the average base quality. The concentration of RNA and INDEX sequence used for library preparation and sequencing is mentioned in Table 1. Raw data obtained was deposited as FASTQ format in NCBI

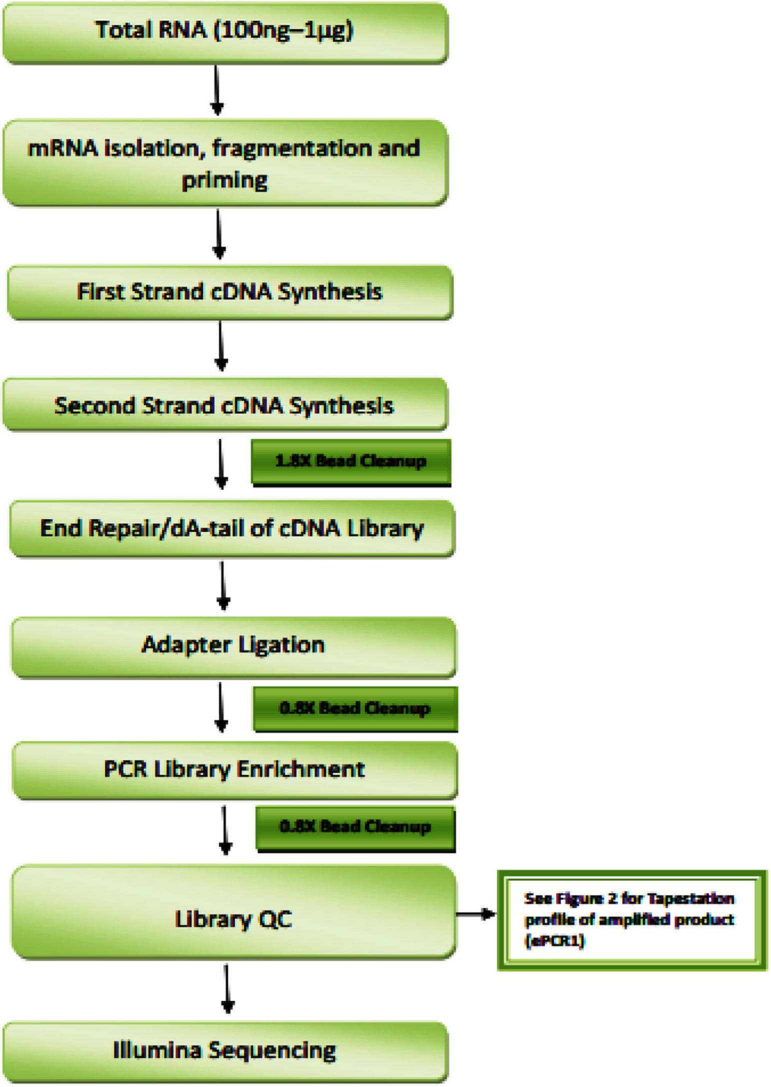
Total RNA (100ng–1µg)
mRNA isolation, fragmentation and priming
First Strand cDNA Synthesis
Second Strand cDNA Synthesis
1.8X Bead Cleanup
End Repair/dA-tail of cDNA Library
Adapter Ligation
0.8X Bead Cleanup
PCR Library Enrichment
0.8X Bead Cleanup
Library QC
See Figure 2 for Tapestation profile of amplified product (ePCR1)
Illumina Sequencing

Fig. 1. Schematic work flow of library preparation and sequencing using NEBNext Ultra Directional RNA Library Preparation kit.

Table 1
Description of the libraries during library preparation and sequencing.

	Sample ID	Qubit (ng/ul)	Vol. (ul)	Qubit Yield (ng)	NEB Barcode	Index Sequences
1	Sonora 64_Control	2.84	10	28.4	1	ATCACG
2	Sonora 64_Treatment	1.99	10	19.9	2	CGATGT
3	Vinata_Control	2.17	10	21.7	3	TTAGGC
4	Vinata_Treatment	4.94	10	49.4	4	TGACCA
5	CRP 1660_Control	2.22	10	22.2	5	ACAGTG
6	CRP 1660_Treatment	3.84	10	38.4	6	GCCAAT
7	DBW 17_Control	6.42	10	64.2	7	CAGATC
8	DBW 17_Treatment	5.01	10	50.1	8	ACTTGA

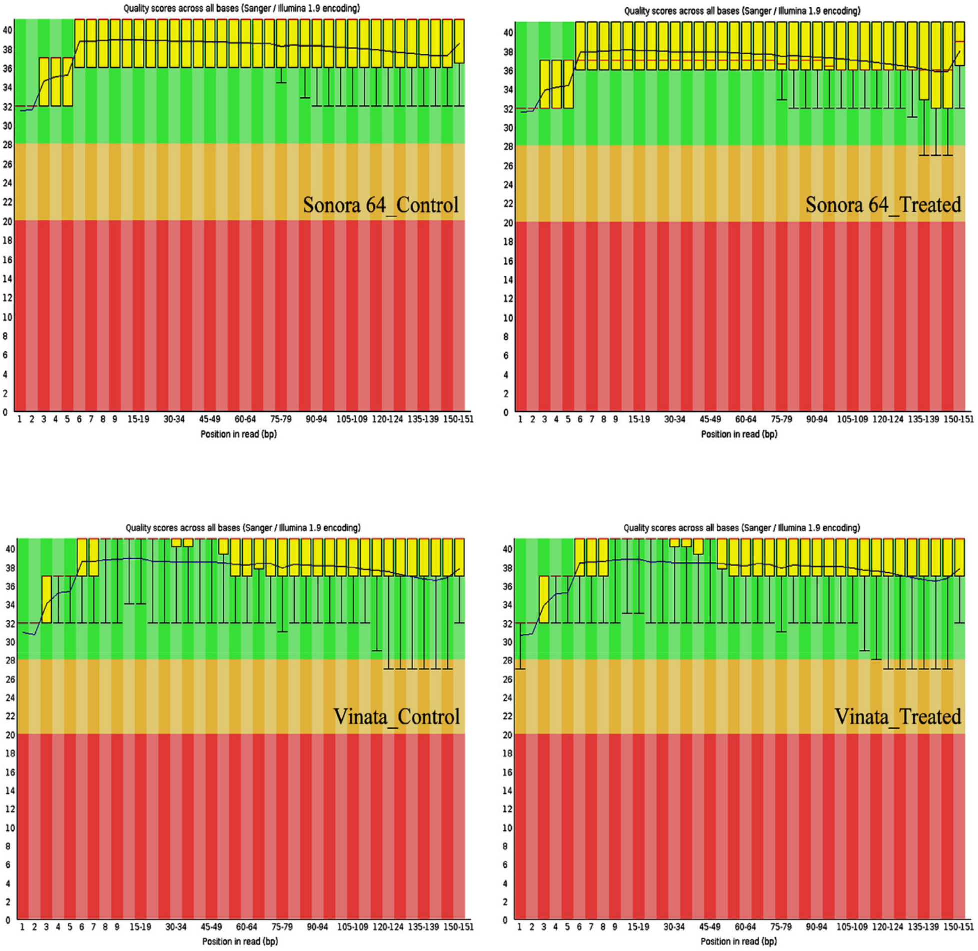

Fig. 2A. Average base quality of reads from Sonora 64 and Vinata genotypes. The position in the read is plotted on the x-axis and the Q-score is plotted on the y-axis. The red line represents the median value of Q-score. The dark blue line is the mean value Q-score. The boxplot represents the inter-quartile range, while the whiskers represent the 10% and 90% points. A Q-score above 30 (>99.9% correct) is considered high quality data.

Sequence Read Archive (SRA) under BioProject # PRJNA605909 as submission # SUB6961770. The accession number for individual wheat variety in NCBI SRA database is given in Table 2. During the assembly, contigs less than 300 bp in length were disqualified as they are too short to depict sequence matches and they may also be deficient in an annotated protein domain resulting in false negatives. Approximately, 95.36% of reads [Fig. 3] were aligned to the clustered transcripts (see experimental design, materials and methods for details) which indicate completeness and quality of the assembly. Out of approximately 47.08 million reads obtained, around ~90.77% of good quality data was utilized for further investigation from each library (Table 3). Around 62.3% of transcripts were characterized against Viridiplantae and 60.91% against *Triticum spp.* (Table 4). Gene Ontology (GO) analysis revealed that biological process (BP) was the major category recognized followed by molecular function and cellular component in all the four assemblies (Fig. 4). Expression trend of transcripts showed that ~70.2%, ~94.2%, ~90.5%,and ~95.2% transcripts in Sonora 64, Vinata, CRP 1660 and DBW 17, respectively, were commonly expressed in both control and treatment samples (Fig. 5). The grain Fe/Zn content analyzed by FAAS [2] and qPCR

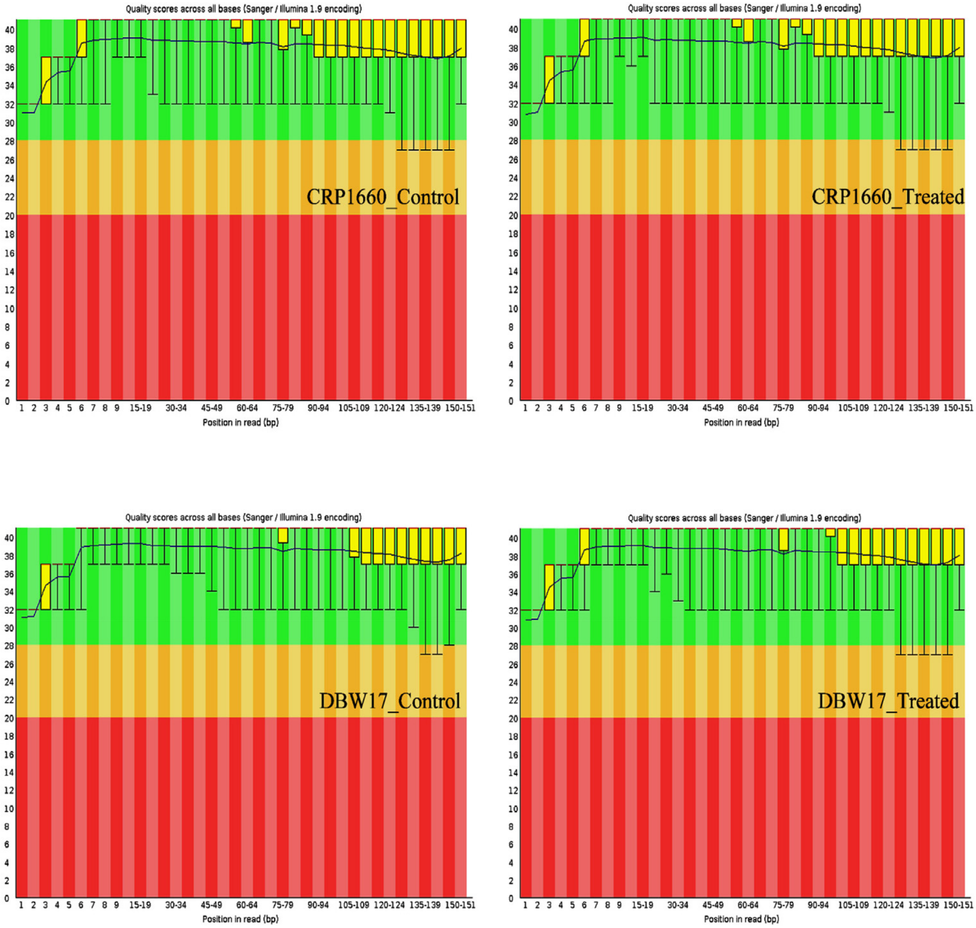

Fig. 2B. Average base quality of reads from CRP 1660 and DBW 17 genotypes. The position in the read is plotted on the x-axis and the Q-score is plotted on the y-axis. The red line represents the median value of Q-score. The dark blue line is the mean value Q-score. The boxplot represents the inter-quartile range, while the whiskers represent the 10% and 90% points. A Q-score above 30 (>99.9% correct) is considered high quality data.

expression data of twelve selected transcripts are shown in supplementary figure S1 and supplementary figure S2, respectively.

Experimental design, materials, and methods

RNA isolation

Frozen whole seedlings were grinded in liquid nitrogen into a fine powder prior to RNA extraction. A sample of 100 mg tissue was utilized for isolating RNA using Qiagen RNeasy Plant Mini kit (Netherland) and using the specified protocol of manufacturer. For analysis of quality and quantity of RNA, $A_{260/280\ \mathrm{nm}}$ readings were taken while for further accurate assessment of sample quality RNA 6000 Nano Assay Kit and Agilent Bioanalyzer 2100 (Agilent, USA) were used for obtaining RIN values. To ensure quality of the final RNA-seq data obtained a threshold RNA integrity number (RIN) reading greater than 8.5 was taken as cut-off for advance operations.

Table 2
List of accession number of individual wheat variety transcriptome in NCBI SRA database.

Variety	Treatment condition	SRA submission accession	Bioproject accession	Biosample accession
Sonora 64_1 (TaxID: 4565)	Control	SUB6961770	PRJNA605909	SAMN14081570
Sonora 64_2 (TaxID: 4565)	Treatment	SUB6961770	PRJNA605909	SAMN14081571
Vinata_1 (TaxID: 4565)	Control	SUB6961770	PRJNA605909	SAMN14081572
Vinata_2 (TaxID: 4565)	Treatment	SUB6961770	PRJNA605909	SAMN14081573
CRP 1660_1 (TaxID: 4565)	Control	SUB6961770	PRJNA605909	SAMN14081574
CRP 1660_2 (TaxID: 4565)	Treatment	SUB6961770	PRJNA605909	SAMN14081575
DBW 17_1 (TaxID: 4565)	Control	SUB6961770	PRJNA605909	SAMN14081576
DBW 17_2 (TaxID: 4565)	Treatment	SUB6961770	PRJNA605909	SAMN14081577

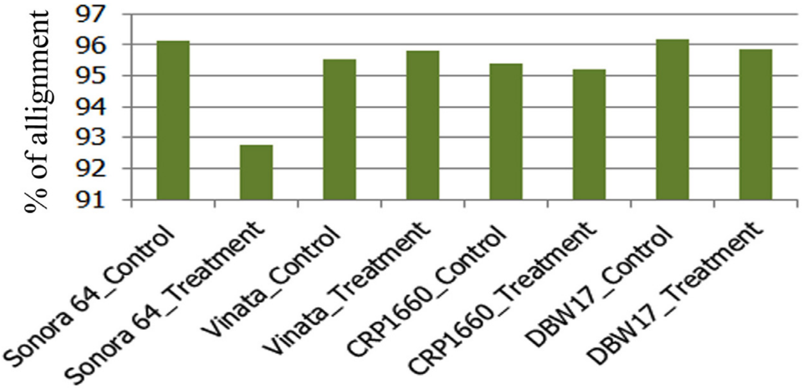

Fig. 3. Read alignment statistics to the clustered transcripts in all the four wheat genotypes. DBW17 control sample shows maximum read alignment (96.2%) to the clustered transcripts with Sonora 64 treated sample being the least (92.77%).

Table 3
Number of raw and processed reads for the samples.

Samples	Raw_Reads	Processed_Reads	% of high quality data
Sonora 64_Control	40,880,754	37,116,391	90.79
Sonora 64_Treatment	41,798,673	38,403,566	91.88
Vinata_Control	44,084,919	39,517,872	89.64
Vinata_Treatment	41,952,459	37,485,278	89.35
CRP 1660_Control	81,636,616	73,789,732	90.39
CRP 1660_Treatment	61,721,999	55,977,866	90.69
DBW 17_Control	53,285,017	49,106,554	92.16
DBW 17_Treatment	49,679,341	45,321,022	91.23

Table 4
Statistics of annotated transcript against Viridiplantae and *Triticum spp.*

Sample name	Total Transcripts	Viridiplantae annotated transcripts	*Triticum* spp. annotated transcripts
Sonora 64	252,177	131,267	125,974
Vinata	189,602	128,977	127,060
CRP 1660	241,791	158,619	155,523
DBW 17	219,813	144,168	141,735
Total	903,383	563,031 (62.3%)	550,292 (60.91%)

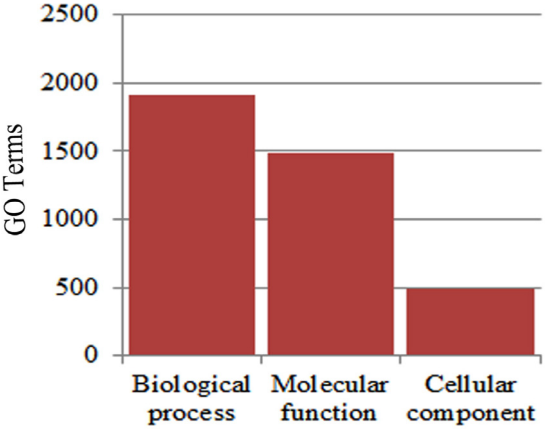

Fig. 4. Graph showing the Gene Ontology terms (y-axis) of the assembled sequences under Biological function, Molecular function and cellular component category (x-axis).

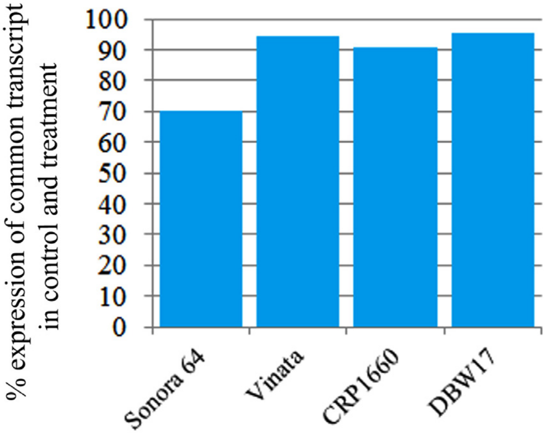

Fig. 5. Graph showing the differential expression analysis (%) (y-axis) of the common transcripts in both control as well as treated samples in all the four wheat genotypes (x-axis).

Library preparation and Illumina HiSeq sequencing

RNA sequencing libraries were generated using Illumina-compatible NEBNext® Ultra^{TM} Directional RNA Library Prep Kit (New England BioLabs, MA, USA) at Genotypic Technology Pvt. Ltd., Bangalore, India. The flowchart ofNEBNext Ultra Directional Library preparation is given in Fig 1. For the isolation of mRNA, fragmentation and priming procedure1 ug of the total RNA was used. First strand synthesis of fragmented and primed mRNA was carried out further with the addition of Actinomycin D (Gibco, life technologies, CA, USA) and second strand synthesis was carried out afterwards. HighPrep magnetic beads (Magbio Genomics Inc, USA) were used for purifying the synthesized double stranded cDNA. Purified cDNA was end-repaired, adenylated and ligated to Illumina multiplex barcode adapters as per NEBNext® Ultra^{TM} Directional RNA Library Prep Kit protocol. Illumina Universal Adapters used in the study were: 5′-AATGATACGGCGACCACCGA GATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′ and Index Adapter: 5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC [INDEX] ATCTCGTATGCCGTCTT CTGCTTG-3′. [INDEX] used here is a distinct sequence for recognizing sequencing data for each sample (Table 1). HighPrep beads were used for purifying adapter ligated cDNA and was subjected to 15 cycles of Indexing-PCR (37 °C for 15 mins followed by denaturation at 98°C for 30 s, cycling (98°C for 10 s, 65°C for 75 s) and 65°C for 5 mins) to enrich the adapter-ligated fragments. Theproduct (sequencing library) obtained finally was subjected to HighPrep beads for purification and later library quality control check was carried out. For the Illumina-compatible sequencing library, fragment size distribution was analyzed using Agilent 2200 Tapestation (Table 1) and quantification by Qubitfluorometer (Thermo Fisher Scientific, MA, USA). Fragment size range between 250 bp to 1000 bp. Since the collective adapter size is approximately 120 bp, effectual specified insert size is 130 to 880 bp which has sufficient concentration to get the desired amount of sequencing data.

Approximately, 51.87 million raw sequencing reads was produced by using Illumina HiSeq sequencer from performing sequencing meant for 150 bp paired-end (PE) reads.

Quality control and processing

Prior to assembly the reads were subjected to processing for assessment of quality and low quality filtering. Generated raw data was assessed for quality by FastQC [3] and preprocessed, encompassing removal of low quality bases (<q30) and the adapter sequences. Cutadapt was used for pre-processing of the data [4]. Overview of average base quality is shown in Fig. 2.

De-novo assembly and sequence clustering

Processed reads were assembled using graph based approach by Trinity [5] program with default k-mer of 25. Assembly performed was genotype specific. This program joins the overlapping reads of a particular quality and length to longer contig sequences devoid of any gaps. Distinguishing characteristics like, including average length, maximum length, N50 length, and minimum length of the arranged contigs were estimated. Clustering of the assembled transcripts on the basis of sequence match is carried out using CD-HIT-EST [6] in the next step of the assembly process among sequences with 95% sequence similarity, which decreases the redundancy without elimination of sequence variety which is used for advanced transcript annotation and the differential expression investigations.

Reads mapping back to transcriptome

The assessment of read content approach was utilised for assessing the quality of the assembly. Bowtie2 [7] with end to end parameters was used for aligning back of processed reads (clustered) to the assembled transcripts from each genotype.

Differential expression analysis

DESeq [8], a R package was utilized for the calculation of differential expression. Size factor estimation in DESeq was used for removal of sequencing (uneven library size/depth) bias among the samples by library normalization.

Annotation

Pfam, KEGG pathway, Uniprot, PlnTFDB and Wheat proteins databases were utilized for the functional annotation of the transcripts. Gene ontology annotation was performed against viridiplantae and wheat protein sequences downloaded from Uniprot server (https://www.uniprot.org/). Annotation of clustered transcripts was carried out by homology approach to allocate functional annotation using BLAST [9] tool against **"**viridiplantae**"** (5716,702) and **"***Triticum* spp." (185,265) proteins from uniprot. If the match was found at minimum similarity greater than 30% and e-value less than e-5 the transcripts were allocated with a homolog protein from another organisms. Metabolic pathway investigation was done by using KAAS [10] Server and identification of pathways was carried out with *Oryza sativa cv. japonica* (Japanese rice) and *Zea mays* (maize) as reference organisms. Simple Sequence Repeats (SSR) were determined for each transcript sequence using MISA [11] perl script. The recommended default parameters of MISA were used for identification of simple repeat of motif length varying from monomer to hexamer. Pfam domain determination was performed using PfamScan in order to understand the conserved domains. Transcripts encoding transcription factors (TF) were determined by homology search against identified plant TFs from PlnTFDB [12].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgments

Authors sincerely acknowledge the financial assistance received from Indian Council of Agricultural Research in the form of project CRP on Wheat Biofortification (Grant No. 1006422). Authors also thank anonymous referees for their constructive comments on the manuscripts. Authors also thank Dr. N.S. Bains and Dr. V.S. Sohu, PAU, Ludhiana India for supplying wheat genotypes CRP 1660.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2020.105995.

References

[1] O.P. Gupta, v. Pandey, R. Saini, S. Narwal, V. Kumar, T. Khandale, S. Ram, G.P. Singh, Identifying transcripts associated with efficient transport and accumulation of Fe and Zn in hexaploid wheat (T. aestivum L.), J. Biotechnol. 316 (2020) 46–55 https://doi.org/10.1016/j.jbiotec.2020.03.015.

[2] V.G. Shobhana, N. Senthil, K. Kalpana, B. Abirami, et al., comparative studies on the iron and zinc contents estimation using atomic absorption spectrophotometer and grain staining techniques (prussian blue and dtz) in maize germplasms, J. Plant Nutri. 36 (2) (2013) 329–342.

[3] FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[4] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet J. 17 (1) (2011) 10–12.

[5] M.G. Grabherr, M.G. Grabherr, B.J. Haas, M. Yassour, et al., Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, Nat. Biotechnol. 29 (7) (2011) 644–652.

[6] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics, (2012), 1;28(23):3150–2.

[7] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359.

[8] S. Anders, W. Huber, Differential expression analysis for sequence count data, Genome Biol. 11 (2010) R106.

[9] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.

[10] Y. Moriya, M. Itoh, S. Okuda, A.C. Yoshizawa, M. Kanehisa, KAAS: an automatic genome annotation and pathway reconstruction server, Nucleic Acids Res. 35 (2007) W182–W185.

[11] S. Beier, T. Thiel, T. Münch, U. Scholz, M. Mascher, MISA-web: a web server for microsatellite prediction, Bioinformat. 33 (2017) 2583–2585.

[12] P.P. Rodriguez, D.M. Riano-Pachon, L.G.G. Correa, S.A. Rensing, et al., PlnTFDB: updated content and new features of the plant transcription factor database, Nucleic Acids Res. 38 (1) (2010) D822–D827.