Journal of Pharmacognosy and Phytochemistry

Available online at www.phytojournal.com

E-ISSN: 2278-4136
P-ISSN: 2349-8234
www.phytojournal.com
JPP 2020; 9(5): 979-986
Received: 21-06-2020
Accepted: 25-07-2020

Madhuri Gupta
Department of Biotechnology, College of Agriculture, Sardar Vallabhbhai Patel University of Agriculture &Technology, Meerut, Uttar Pradesh, India

Pushpendra Kumar
Department of Biotechnology, College of Agriculture, Sardar Vallabhbhai Patel University of Agriculture &Technology, Meerut, Uttar Pradesh, India

Jitender Singh
College of Biotechnology, Sardar Vallabhbhai Patel University of Agriculture & Technology, Meerut, Uttar Pradesh, India

Devendra Kumar
ICAR-Central Potato Research Institute Campus, Modipuram, Meerut, Uttar Pradesh, India

Anil Sirohi
Department of MBGE, College of Biotechnology, SVPUAT, Meerut, Uttar Pradesh, India

Mukesh Kumar
Department of Genetics & Plant Breeding, College of Agriculture, Sardar Vallabhbhai Patel University of Agriculture &Technology, Meerut, Uttar Pradesh, India

Mukesh Kumar
Department of Biotechnology, College of Agriculture, Sardar Vallabhbhai Patel University of Agriculture &Technology, Meerut, Uttar Pradesh, India

Corresponding Author:
Madhuri Gupta
Department of Biotechnology, College of Agriculture, Sardar Vallabhbhai Patel University of Agriculture &Technology, Meerut, Uttar Pradesh, India

Computational analysis of potato (*Solanum tuberosum*) transcriptomic RNA-seq data quality using CLC workbench

Madhuri Gupta, Pushpendra Kumar, Jitender Singh, Devendra Kumar, Anil Sirohi, Mukesh Kumar and Mukesh Kumar

Abstract
The CLC *de novo* assembler is a support system for large data sets and integrated scaffolding for joining contigs based on paired reads information. It is designed to accept a combination of data from Illumina, 454, SOLiD, Ion Torrent and Sanger sequencing as a mix of paired and unpaired reads which allows the quality analysis of SRA data from different sequencing technologies, to be exploited. In the present study, transcriptome data of three different physiological stages of potato tubers i.e. Dormant tuber (DT), Dormancy release tuber (DRT) and Sprouting tuber (ST) were studied for computational quality analysis. A total no. of 38,714,870 paired sequences of DT (ID: SRR1039535), 38,669,102 paired sequences of DRT (ID: SRR1103933) and 38,753,150 paired sequences of ST (ID: SRR1103934) were downloaded from SRA database in FastQ format. The *De novo* assembly resulted with 50849 (DT), 44550 (DRT) and 46254 (ST) contigs. The values for N50 and average contig length recorded 1,016(DT), 1,118 (DRT), 1,015 (ST) and 680(DT), 725(DRT), 685(ST) respectively They have been analysed for quality check with the *Per-sequence analysis, Per-base* and *Over-representation analysis* using CLC workbench which includes different categories resulting in good quality of data.

Keywords: RNA-seq, *Solanum tuberosum*, CLC

Introduction
Sequencing technologies are in continuous development with new technologies emerging, improvements in the sequencing quality and rapidly increasing amounts of sequencing data. Combined with the fact that we for the last couple of years have seen a significant increase in the number of researchers involved in next generation sequencing projects creates a need for *de novo* assemblers for handling extremely large amounts of data and solve the complex task of constructing de novo assemblies from short read data (Chinnappa and McCurdy, 2015) [2]. Biologists and sequencing facility technicians face not only issues of minor relevance, e.g. suboptimal library preparation, but also serious incidents, including sample contamination or even mix-up, ultimately threatening the accuracy of biological conclusions. Unfortunately, most of the problems and evolving questions raised above can't be solved and answered entirely. There are large number of short read assemblers available with their own strength and weaknesses *viz.*, DNASTAR, Spades, SOAP-denovo, MIRA, ALLPATHS, CLC Workbench etc. However, the sequencing data quality control tool of the CLC Genomics Workbench provides various generic tools to assist in the quality control process of the samples by assessing and visualizing statistics based on Sequence-read lengths and base-coverages, Nucleotide-contributions and base-ambiguities, Quality scores as emitted by the variation detection tool, Over-represented sequences and hints suggesting contamination events.

To understand the various mechanisms taking place at different stages, genomics and transcriptome analysis is pivotal in present biological scenario. Tuber Dormancy is very important phenomenon and its break is known as "Endodormancy break" which leads to early and prominent sprouting of potato field crop. The detailed study of this mechanism using available pooled transcriptomic data of three different stages of potato tuber *viz.*, DT (Dormant Tuber), DRT (Dormancy release tuber) and ST (Sprouting Tuber) has been downloaded from publicly available SRA database from NCBI, which has been developed through next-generation sequencing (NGS) technology ILLUMINA (Illumina HiSeq 2000) (Liu, *et al.*, 2015) [9]. The *De novo* assembly of these sequence reads was done using evolved software CLC Genomics Workbench to generate the contigs which resulted in 50849 (DT), 44550 (DRT) and 46254 (ST) contigs. Further, the quality check has been performed of RNAseq data with parameters of *Per-sequence analysis, Per-base* and *Over-representation analysis* which includes different components for the assurance of the data to be of good quality.

Materials and methods

Plant sample details and SRA data downloading

The dataset of transcriptomic reads of three different potato tuber stages *viz.*, Transcriptome of DT (Dormant tuber, SRR1039535), DRT (Dormancy release tuber, SRR1103933) and ST (Sprouting tuber, SRR1103934) was downloaded from ENA in FastQ format. These reads were developed using RNA isolation and freezing in liquid nitrogen by collecting the tubers after harvest at each time point *viz.*, 0 day (Dormant tuber), 30 day (Dormancy release array start after harvest at room temperature) and the tubers with 2-3mm length buds were defined as sprouting tubers.

Three stages are being used to do the comparative analysis obtained through Paired-end sequencing on an IlluminaHiSeqTM2000 (7.74 GB (DT), 8.54 GB (DRT) and 8.56 GB (ST)) by Liu *et al.*, 2015 [9].

System requirements and de-novo assembly

The system requirements of CLC Main Workbench includes Windows 7/8/10, Windows Server 2012, Windows Server 2016 and Windows Server 2019, OS X 10.10, 10.11, 64-bit operating system, 2 GB RAM recommended and a minimum of 1024 x 768 display is required.

Downloaded paired SRA reads in FASTQ format were analyzed to retrieve information about read length distribution, GC content, nucleotide base ambiguity, sequence quality, sequence duplication levels etc using CLC genomics workbench (www.qiagenbioinformatics.com/products/clc-genomics-workbench/). Keeping statistical parameters as default, trimmed reads were used as input file for further *de-novo* assembly using CLC Genomics workbench and reads were assembled to form contiguous consensus sequences (contigs) from collections of overlapping reads in fasta format.

Results

The SRA data downloaded from NCBI has been analysed using CLC Genomics Workbench which habours various generic tools to assist in the quality control process of the samples. It works by assessing and visualizing statistics of Sequence-read lengths and base-coverages, Nucleotide-contributions and base-ambiguities, Quality scores as emitted by the variation detection tool, Over-represented sequences depicting contamination events.

De novo assembly

De novo assembly of potato reads has been carried out to generate contig sequences with CLC genomics workbench and contigs were formed which includes the scaffolded regions (Table 1). Scaffolds are the linking together of non-contiguous series of genomic sequences which are first assembled into contigs which by nature of their assembly have gaps between them and to bridge the gaps between these contigs was considered to create a scaffold. In table 1, the RNA-seq reads from a previous study NCBI Short Read Archive Bioproject for Dormant Tuber (ID: SRR1039535), Dormancy Release Tuber (ID: SRR1103933) and Sprouting Tuber (ID: SRR1103934) were downloaded from SRA database (Liu *et al.*, 2015) [9]. The computational analysis study and assembly performed using CLC Genomics Workbench, recorded were 50849 (DT), 44550 (DRT) and 46254 (ST) contigs (Table 1). The values for N50 and average contig length recorded 1,016(DT), 1,118 (DRT), 1,015 (ST) and 680(DT), 725(DRT), 685(ST) respectively.

Table 1: Length measurement of DT, DRT and ST contigs. N75 is a length of contigs required to cover 75% of total transcriptome (similarly for N50 and N25).

	Contig length (including scaffold)		
	DT	**DRT**	**ST**
N75	456	503	467
N50	1,016	1,118	1,015
N25	1,852	1,907	1,794
Minimum	200	200	200
Maximum	10,531	10,154	11,459
Average	680	725	685
Count	50,849	44,550	46,254
Total	34,598,951	32,277,592	31,668,114

These computational parameters explained the quality of data on the basis of per-sequence, per-base and over-representation analysis which are presented as follows:

Per-sequence analysis: To check the quality on the basis of per-sequence, the parameters taken in account include distribution of sequence lengths, distribution of GC-contents, distribution of N-contents and distribution of average sequence quality.

1) **Distribution of sequence lengths:** It reveals the number of sequences that have been observed for individual sequence lengths. X-axis is denoting the sequence length and Y-axis is explaining percentage of sequences (% sequences). The resulting graphs correlates sequence-lengths in base-pairs with numbers of sequences observed with that number of base-pairs (figure 1). It has been observed that for all the three data figure 1(a), (b), (c), the number of bases that make up the sequence length is 90 and no secondary peaks at unexpected lengths have been observed to be trimmed.

2) **Distribution of GC-contents** shows the counts in the number of sequences that feature individual % GC-contents ranging from 0 to 100%. In Fig.2, the X-axis is explaining the percentage of GC content (% GC Content) and Y-axis as percentage of sequences (% sequences). The % GC-content of a sequence is been calculated by dividing the absolute number of G/C-nucleotides by the length of that sequence. The pattern in all the three data showed normal distribution in the range of 20-70 GC% (figure 2 (a), (b), (c)) indicating good quality SRA data to be used further.

3) **Distribution of N-contents** counts the number of sequences that feature individual %N-contents from 0 to 100%, where N refers to all ambiguous base-codes as specified by IUPAC. The X-axis defines the percentage of ambiguous bases (% ambiguous) while Y-axis determines the percentage of sequences (% sequences). The % N-content of a sequence is calculated by dividing the absolute number of ambiguous nucleotides through the length of that sequence. Ambiguous nucleotide distribution is close to 0 as pattern (figure 3(a), (b), (c)). Thus, this demonstrated that the SRA data has good quality.

4) **Distribution of average sequence quality** scores calculates the amount of sequences that feature individual PHRED-scores from 0 to 63. The quality score of a sequence as calculated as arithmetic mean of its base qualities. For Quality distribution graph, X-axis determines the average PHRED-Score and Y-axis as

percentage of sequences (% sequences). PHRED-scores of 30 and above are considered high quality and they are observed in the same pattern ranging between 25-40 (figure 4 (a), (b), (c)). So, the data was considered to be of high quality and was further analysed for more biological analysis.

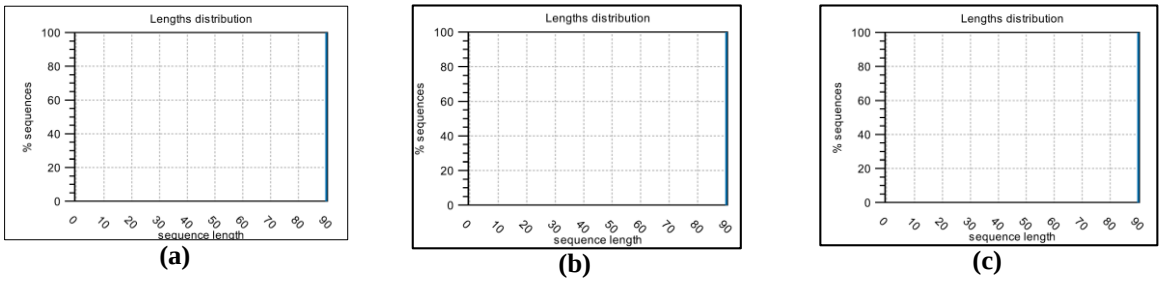

Fig 1: Distribution of sequence lengths **x:** sequence length in base-pairs **y:** number of sequences featuring a particular length normalized to the total number of sequences DT (a), DRT (b), ST (c).

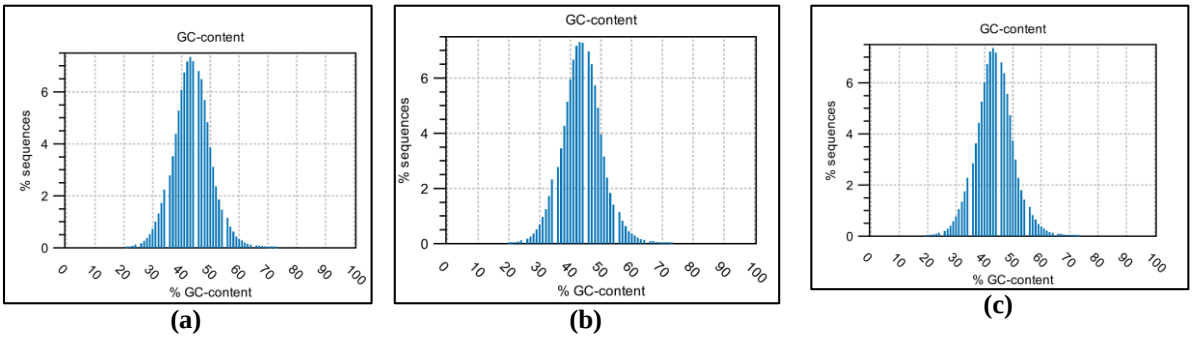

Fig 2: Distribution of GC-contents. The GC-content of a sequence is calculated as the number of G C-bases compared to all bases (including ambiguous bases). **x:** relative GC-content of a sequence in percent, **y:** number of sequences featuring particular GC-percentages normalized to the total number of sequences. DT (a), DRT (b), ST (c).

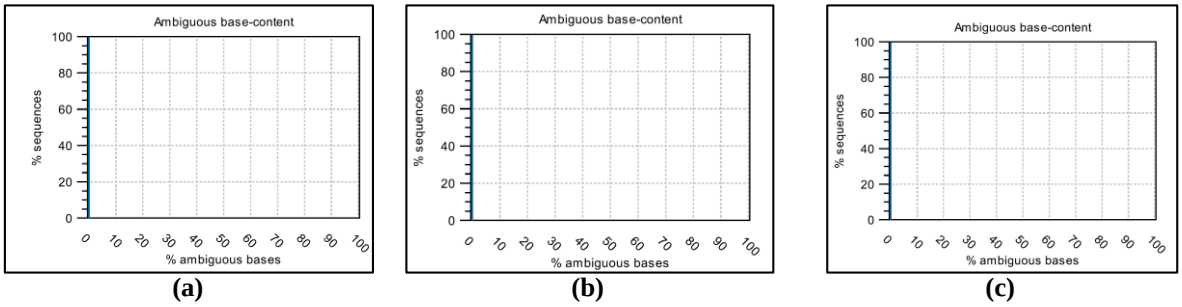

Fig 3: Distribution of N-contents. The N-content of a sequence is calculated as the number of ambiguous bases compared to all bases. **x:** relative N-content of a sequence in percent, **y:** number of sequences featuring particular N-percentages normalized to the total number of sequences. DT (a), DRT (b), ST (c).

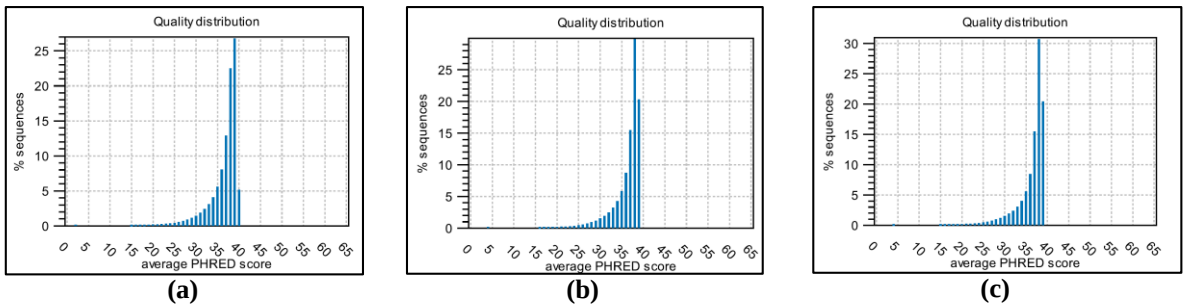

Fig 4: Distribution of average sequence qualitie scores. The quality of a sequence is calculated as the arithmetic mean of its base qualities. **x:** PHRED-score, **y:** number of sequences observed at that qual. score normalized to the total number of sequences. DT (a), DRT (b), ST (c).

Per-base analysis: Quality check takes into account the data on the basis of per-base parameters for Coverages for the four DNA nucleotides and ambiguous bases, Combined coverage of G- and C-bases and Base-quality distribution.

1) **Coverages for the four DNA nucleotides and ambiguous bases** calculates absolute coverages for the four DNA nucleotides (A, C, G or T) for each base position in the sequences. In graph 5 (a), (b) and (c), the x-axis is showing the base positions and y-axis is showing the percentage of nucleotide contributions and ambiguous bases for all the three datasets in a random library. There was little or no difference between the

bases, thus the lines in this plot are parallel to each other and as observed Adenine and Thymine are parallel to Guanine and Cytosine, whereas, no nucleotide can be seen as Ambiguous. The same pattern is observed in all the three data which indicates good quality of the data.

2) **Combined coverage of G- and C-bases** as shown in figure 6 (a), (b), (c) calculates absolute coverages of C's + G's for each base position in the sequences. X-axis is showing base position and y-axis is showing the percentage of GC content and all the three data are showing the range of 40-80 percent of GC content and no GC biasness observed with changes at specific base positions along the read length. So, this could not indicate that an over-represented sequence is contaminating the library. The pattern obtained showed the expected results and a uniform pattern was observed in all the three data.

3) **Base-quality distribution** along the base positions calculates the number of bases that feature individual PHRED-scores in 64 bins from 0 to 63. X-axis is showing base position and y-axis is showing the PHRED Score and in all the three data. We can observe that contigs are lying in between 75-95 percentile region in a uniform pattern and approximately 5 percentile of data is going below the score of 20 which can be considered very low and can be neglected. PHRED-scores above 20 are considered as good quality as being mentioned in the user manual of CLC workbench and according to the statistics, it is normal to find the quality dropping off near the end of reads. This could be observed in all the three data's as shown in figure 7 (a), (b) and (c).

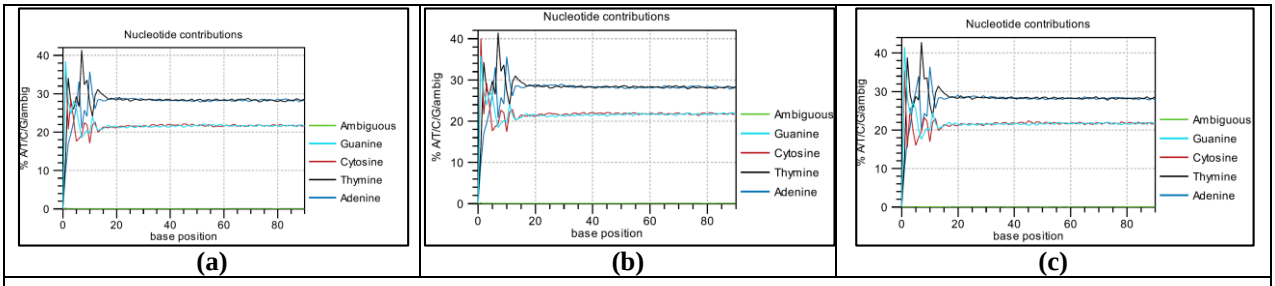

Fig 5: Coverages for the four DNA nucleotides and ambiguous bases: **x**: base position, **y:** number of nucleotides observed per type normalized to the total number of nucleotides observed at that position. DT (a), DRT (b), ST (c).

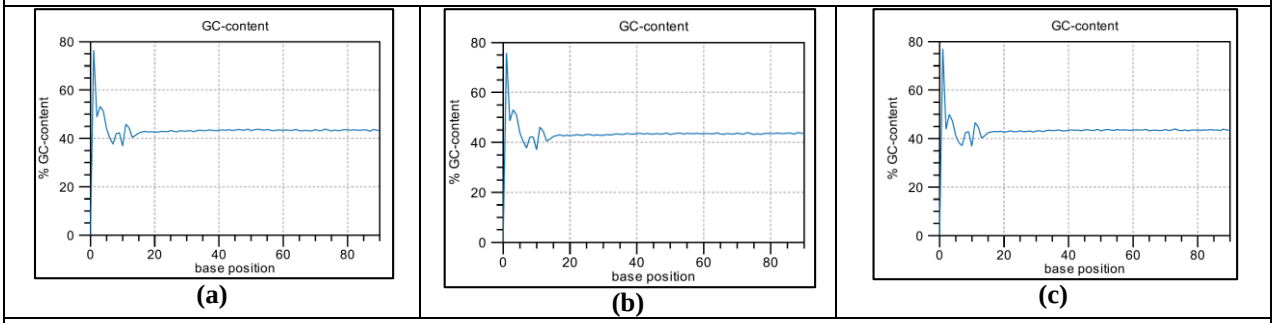

Fig 6: Combined coverage of G- and C-bases: x: base position, y: number of G- and C-bases observed at current position normalized to the total number of bases observed at that position. DT (a), DRT (b), ST (c).

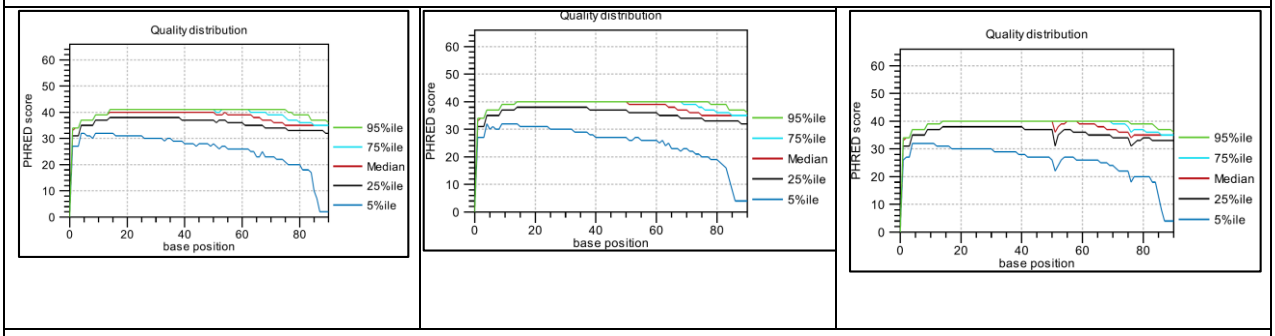

Fig 7: Base-quality distribution along the base positions: x: base position, y: median & percentiles of quality scores observed at that base position. DT (a), DRT (b), ST (c).

Over-representation analysis: The third parameter taken into account for quality analysis was over-representation analysis which explains the five most-overrepresented 5mers, duplication level distribution and the duplicated sequences present in the RNA-seq data.

1) **The five most-overrepresented 5mers**: The 5-mer analysis examines the enrichment of penta-nucleotides. The enrichment of 5-mers is calculated as the ratio of observed and expected 5-mer frequencies. X-axis describes the base position while on Y-axis the percentage coverage is being plotted. In all the three data the top five enriched 5-mers were observed. TCTTC is the most enriched 5-mer in Dormant tuber (DT) while CTTCT is in Dormancy release Tuber (DRT) and Sprouting Tuber (ST) stage followed by CTTCT in DT

and TCTTC in DRT and ST as shown in figure 8(a), (b), (c).

2) **Duplication level distribution** identifies sequence reads that have been sequenced multiple times. Duplication levels are simply the count of how often a particular sequence has been found. X-axis is for duplicate count and Y-axis for relative sequence count. In fig. 9 (a) and 9 (c), it was observed that the relative sequence count was reaching to a value of 40 and 38 while in 9 (b) it has been recorded near 33. A low level of duplication indicates no enrichment bias introduced by PCR amplification (figure 9 (a), (b), (c)).

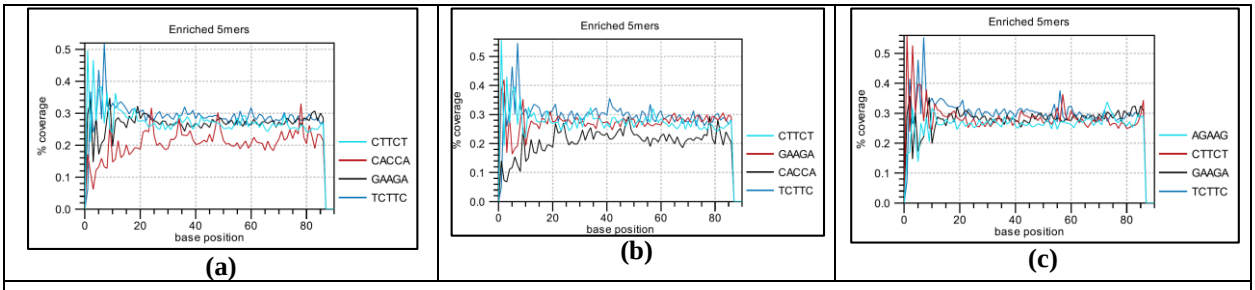

Fig 8: The five most-overrepresented 5mers. The over-representation of a 5mer is calculated as the ratio of the observed and expected 5mer frequency. The expected frequency is calculated as product of the empirical nucleotide probabilities that make up the 5mer. (5mers that contain ambiguous bases are ignored): x: base position, y: number of times a 5mer has been observed normalized to all 5mers observed at that posit ion. DT (a), DRT (b), ST (c).

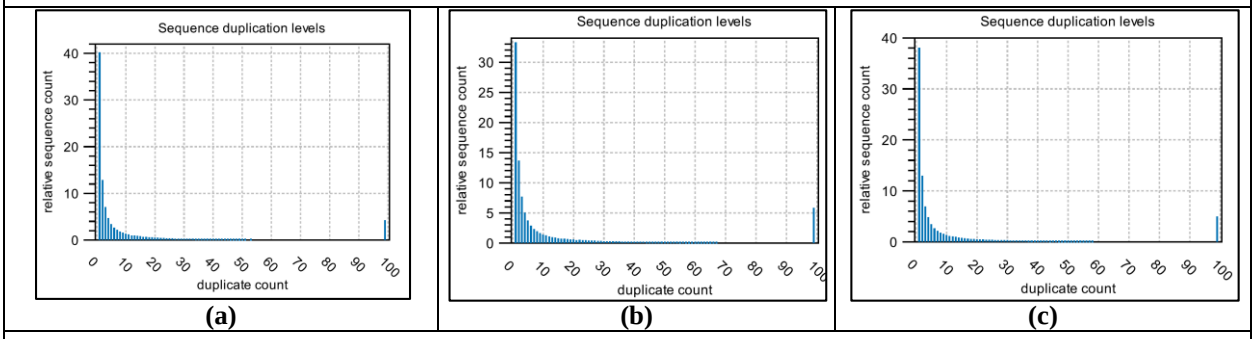

Fig 9: Duplication level distribution: x: duplicate count, y: number of sequences that have been found that many times normalized to the number of unique sequences. DT (a), DRT (b), ST (c).

3) **Duplicated sequences** are the over-represented sequences that are reported more than once. The table showing the sequences, their number of times to be observed and normalized to the number of unique sequences are given below in Table 2 in Dormant Tuber (DT), Dormancy release tuber (DRT), Sprouting Tuber (ST) respectively. In, Duplicated sequences, there are 09 sequences found to be common in all of the three data, 07 are found common in two datasets while 09, 10 and 13 duplicated sequences are found exclusively for Dormant tuber, Dormancy release tuber and sprouting tuber respectively. The parameters explained in tables are given below:

- **Sequence**: the 5'-end of the sequence that has been found multiple times
- **abs:** number of times this sequence has been observed
- **%:** number of times this sequence has been observed normalized to the number of unique sequences.

Table 2: Duplicated sequences for common for DT, DRT and ST

S. No.	Sequence (DT)	abs			%		
		DT	DRT	ST	DT	DRT	ST
1)	CGATCTTTCACTTCTGAGAATCCA ATTGTCCTCCCCACAACTTGTCATGA	21,721	8,313	7,758	0.10	0.04	0.04
2)	CTTTGCTCGATCTTTCACTTCTGA GAATCCAATTGTCCTCCCCACAACTT	17,136	8,982	17,521	0.08	0.04	0.04
3)	GCCTTGATGAACACAAATGGAATA AAGTCCTTATCGTCATCGACAGTCAC	16,241	7,663	7,663	0.07	0.04	0.04
4)	GCTCGATCTTTCACTTCTGAGAAT CCAATTGTCCTCCCCACAACTTGTCA	14,271	6,276	7,319	0.07	0.03	0.03
5)	GTCGATTAATATTTTGAGTTTCCTC TTGCTTTCAAGTACCCTCTCTTTGG	13,895	6,777	9,064	0.06	0.05	0.05
6)	TGGGCATTGAAGGTTTCCAATATT GTCCAAGTATACGGCTCCGGCCCCGA	10,187	5,537	11,161	0.05	0.03	0.05
7)	GTTTATTTTAATTTACAACATAACA TATATAGTAGCTGCTAAAACATATA	8,587	7,369	10,131	0.04	0.03	0.05
8)	CTCTTTACAACTAAAAGAAAATGG AGTTAAAGTTTGCTCACATCATTGTT	8,211	26,086	10,906	0.04	0.12	0.05
9)	CCCTTAGCAAGCTTTGTTGGTACA	8,126	22,039	7,691	0.04	0.10	0.04

	CCAATAAGTTCTGGCCAGCTTAGTTT						
10)	CGCCGATATTTTTGCACTGAAGAT GAGAGGGACAATGCAGTAACTTGTAT	9,619	6,038	-	0.04	0.03	-
11)	GGGGCATCTGAATTTGGGGACTTA CCTAGGTACACATCACCACCTAACGC	9,299	5,612	-	0.04	0.03	-
12)	GGCTTACGGTGGATACCTAGGCA CCCAGAGACGAGGAAGGGCGTAGTAAT	8,680	11,551	-	0.04	0.05	-
13)	GCTTTCTTTTCCTCTGGCTACTAA GATGTTTCAGTTCGCCAGGTTGTCTC	8,645	11,275	-	0.04	0.05	-
14)	TGGGCATTGAAGGTTTCCAATATT GTACAAGTATACGGCTCCGGCCCCGA	10,927	-	8,486	0.05	-	0.04
15)	GTTTGGGCATTGAAGGTTTCCAAT ATTGTCCAAGTATACGGCTCCGGCCC	10,416		11,210	0.05	-	0.05
16)	GGCCAATCCAGAAGATGGACAAG TCTAGGGTCACATTGCAGGGTACATAT	11,814	-	-	0.05	-	-
17)	GTCGTTTTCATTTCCTACCTTACCA CCAGTTACCACCAACTGTTCATCAT	8,929	-	-	0.04	-	-
18)	GCCTTGATGAACACAAATGGAATA AAGTCCTTTTCGTCATTGACAGTCAC	8,331	-	-	0.04	-	-
19)	CGTTGATATTGTTCAAGGAAACGG GGAGCATTCACTGTATGTACCACCAC	8,178	-	-	0.04	-	-
20)	GCCGCCGATATTTTTGCACTGAAG ATGAGAGGGACAATGCAGTAACTTGT	7,216	-	-	0.04	-	-
21)	GAAAAAAAGGCAGTACTAATTAAT TATCCATCATGGCTGTTCACAAGGAA	6,955	-	-	0.03	-	-
22)	GTTGTCTTGAGGTGCACCACTAAT ACCAGCAGGAGGGTCCTGCTGCAACC	6,788	-	-	0.03	-	-
23)	CGGGGAGCATTCACTGTATGTACC ACCACTGGGGGACTTCTTCAGCTTAC	6,369	-	-	0.03	-	-
24)	GCTCGATTCAGTTATTTCTACTTGT TTCGCTGCCTTTGGTTTGAAAATCA	6,302	-	-	0.03	-	-
25)	CCCCTTAGCAAGCTTTGTTGGTAC ACCAATAAGTTCTGGCCAGCTTAGTT	-	14,763	-	-	0.07	-
26)	CTTTACAACTAAAAGAAAATGGAG TTAAAGTTTGCTCACATCATTGTTTT	-	12,061	-	-	0.06	-
27)	CTACTCTTTACAACTAAAAGAAAAT GGAGTTAAAGTTTGCTCACATCATT	-	8,176	-	-	0.04	-
28)	GCCTGTTAATGCATTTTCTTGAAC CCTGAGGTAATTGTTTTGTGAATGAC	-	7,321	-	-	0.03	-
29)	GTTTTCTCTGCCCTCCTCAACCCT AAAAAACCGAAAAATTCTCTCTTCAG	-	7,138	-	-	0.03	-
30)	CCTTAGCAAGCTTTGTTGGTACAC CAATAAGTTCTGGCCAGCTTAGTTTT	-	6,942	-	-	0.03	-
31)	GGGCCATGTTCGAAGTAAAAGGG TACAATATCTTTGGCAGCAGCAAAGGG	-	6,537	-	-	0.03	-
32)	CTTTCTTTTCCTCTGGCTACTAAGA TGTTTCAGTTCGCCAGGTTGTCTC T	-	6,413	-	-	0.03	-
33)	GGCTTATTTGTTTTGATGTCAAAG CTTGAGATGGCAACTTCTGTCAAAGC	-	6,079	-	-	0.03	-
34)	CCCGGGCATTGAGAAGGAAGGAC GCTTTCAGAGGCGAAAGGCCATGGGGA	-	5,954	-	-	0.03	-
35)	GTGCATGTTCTGGATCTTTCTTTG CCTCATGTTTCTCATGCAAGGCGTAG	-	-	9,467	-	-	0.05
36)	GTTTGTGTGCATGTTCTGGATCTT TCTTTGCCTCATGTTTCTCATGCAAG	-	-	9,377	-	-	0.04
37)	CTCTCTTTGGTTGCCTTTGCTCGA TCTTTCACTTCTGAGAATCCAATTGT	-	-	9,231	-	-	0.04
38)	CTTTGGTTGCCTTTGCTCGATCTT TCACTTCTGAGAATCCAATTGTCCTC	-	-	8,985	-	-	0.04
39)	CTTTCTCCATAGTCATAAGAATCG ATCTGATGGAGGAGGAGAAACACCAC	-	-	8,426	-	-	0.04
40)	GTGTGTACAAAGGGCAGGGACGT AGTCAACGCGAGCTGATGACTCGCGCT	-	-	8,414	-	-	0.04
41)	TAAGAATCGATCTGATGGAGGAG GAGAAACACCACCACCACCTGTTCCAC	-	-	8,230	-	-	0.04
42)	GGGCAACATTCATCAATTGAAGTT GTGTTGAACATGCAAGTCTTCACTTT	-	-	8,087	-	-	0.04
43)	TGCATGTTCTGGATCTTTCTTTGC CTCATGTTTCTCATGCAAGGCGTAGG	-	-	7,895	-	-	0.04
44)	CAAAAACTTACTATCCTTTTCACCA TTCTCCTTGTGGTTATTGCTGCTCA	-	-	7,534	-	-	0.04

45)	GGAGCATTCACTGTATGTACCACC ACTGGGGGACTTCTTCAGCTTACACC	-	-	7,222	-	-	0.03
46)	CTCGATCTTTCACTTCTGAGAATC CAATTGTCCTCCCCACAACTTGTCAT	-	-	6,923	-	-	0.03
47)	CTTAGATGTTCTGGGCCGCACGC GCGCTACACTGATGTATTCAACGAGCT	-	-	6,763	-	-	0.03
48)	GTCGATTGTTTTCATTTGGAGTAG TTATCATAGCAGTCAATAAACCTCCT	-	7,008	9,086	-	0.03	0.04

Discussion

With the accumulation of experimental transcriptomic RNA seq data in the field of plant sciences along with the development of new tools for bioinformatic analysis, it has become possible for multi-omics to jointly analyze the certain life phenomenon (Zhang *et al.*, 2010; Lakshmanan *et al.*, 2015) [13, 10]. SRA data quality analysis can also provide an insight into the data repository itself. Basic quality values, are important to obtain an overview of the archive, which illustrate the overall distribution of data and its quality (Ohta *et al.*, 2017) [11]. Thus, the transcriptomic data of potato (*Solanum tuberosum*) tuber at different stages, DT, DRT and ST from SRA Database (Liu *et al.*, 2015) [9] derived from Illumina/Solexa sequencing technology is used to check the quality through computational analysis (CLC, 2015) [3].

In present study, CLC Workbench has been used for *De novo* assembly, which generated 50849 (DT), 44550 (DRT) and 46254 (ST) contigs and the N50 values recorded 1,016(DT), 1,118 (DRT), 1,015 (ST). The *De novo* RNA seq assembly of legume *Vicia sativa* L. by CLC Genomics Workbench also reported 22748 contigs with N50 of 588 bp (Hetalkumar, 2015) [7]. Similarly, the data analysed from Leaf, root, and flower tissues of Red clover plant using CLC workbench also reported that the denovo assembly resulted in 37,565 contigs with N50 value of 1707 (Chakrabarti *et al.*, 2016) [1]. Celery tissues have been reported to have 42,280 unigenes with an average length of 502.6 bp and an average length of 604 bp (Fu *et. al.*, 2013) [5]. Further, computational analysis of SRA data using publically available programs, Velvet (v1.2.07), Oases (v0.2.08), ABySS (v1.2.7) and commercially available CLC Genomics workbench (v4.7.2) for *de novo* assembly, has reported that CLC workbench is showing the significant results with different parameters for the quality check of data and also resulting in better number of contigs as compared to the other available programmes (Kotwal *et al.*, 2016) [8].

We focussed on the quality assurance and sample authenticity parameters by taking into account three major components: per-sequence, per-base and over-representation analysis. Per-sequence based analysis reveals firstly the distribution of sequence lengths which is found to be optimum that make up the sequence length and no secondary peaks at unexpected lengths have been observed to be trimmed. Secondly, the distribution of GC-contents is found in a range of 20-70 GC% which indicates a good quality data. The quality analysis of longan (*Dimocarpous longan Lour.*) transcriptome has also been performed and recorded the similar GC percentage within the same range using CLC Workbench (Goyal *et al.*, 2017) [6]. Similarly, Roy *et al.*, 2018 [12] has done computational analysis on Korean Medicinal Herb (*Cirsium japonicum)* which is a medicinal plant in Asia and reported the GC content distribution within 42%–45%. Thirdly, the Distribution of N-contents and average sequence quality has been examined showing the best-known results, as close to 0 as possible and PHRED-scores of 25-40 respectively is again an indication of high-quality data. As per the conventions adopted by the Open Bioinformatics Foundation for quality score, the Illumina FASTQ variant encodes PHRED scores which can hold 0-62 PHRED scores (Cock *et al.*, 2009) [4].

In per-base analysis, the coverages for the four DNA nucleotides and ambiguous bases depicts the parallel view between the complementary bases and no ambiguous nucleotide has been found. In Combined coverage of G- and C-bases, a percent range 40-80 of GC content with no GC biasness or changes at specific base positions is observed. So, no contamination of an over-represented sequence could be indicated in the data. The pattern is showing the expected results and a uniform pattern is observed in all the three data. In the third parameter of this category, Base-quality the contigs are lying in the 75-95 percentile region in a uniform pattern and approximately 5 percentiles of data in all the three categories is dropping off near the ends which is normal according to the statistics present in the user manual of CLC Workbench.

The RNA-seq data retrieved and analysed further for over-representation analysis in which the five most-overrepresented 5mers has been observed for three different growth stages where they are found to be different in Dormant stage when compared to Dormancy release and Sprouting stage of the potato tuber. It shows that the different molecular phenomenon and difference in the expression of genes is taking place at the most transition phase of Dormant to Dormancy release stage leading to the difference in pentamers. Similarly, Goyal *et al.*, 2017 [6] also reported the over representation of 4 identified pentamers (5mers) CACCA, TGGTG, CCACC, GGTGG which were calculated as the ratio of the observed and expected 5mer frequency in longan transcriptome quality analysis. In Duplication level distribution, a low level of duplication indicates no enrichment bias introduced by PCR amplification. In, Duplicated sequences, there are 09 sequences found to be common in all of the three data, 07 are found common in two datasets while 09, 10 and 13 duplicated sequences are found exclusively for Dormant tuber, Dormancy release tuber and sprouting tuber respectively. These sequences are found to be duplicated in terms of the 5'end of the sequence that has been found multiple times in the data.

All the parameters discussed above when taken into consideration using CLC Workbench software for computational analysis of the RNA-seq data, indicates the compatibility and possibility of reuse of data for data quality analysis. The high-quality data can be further used for annotation and gene expression analysis, to explore the complicated vital phenomenon of plant life cycle especially dormancy break in potato related tubers.

Conclusion

CLC Workbench is very much user-friendly software which can be easily used by the common biologists even if not having much knowledge of system biology for reanalysing the available open-source RNA-seq data. The potential need of RNA-seq data is to firstly be quality assured which is the major key interest of any biotechnologist for further analysis and extracting results from any data. The reanalysis of open

source data has tremendous potential to work out on various aspects of molecular biology at gene and transcription factors level, working on the biological pathways, depicting the enzyme classes and sub-classes with their vital roles, study on the differential gene expressions etc. So, the computational analysis of RNAseq data for quality fulfils all the required fields needed for the quality assurance and can now be proceeded for functional analysis.

Acknowledgement
We sincerely acknowledge Dr. Huaijun Si, College of Life Science and Technology, Gansu Agricultural University, Lanzhou, Republic of China for permit to use their SRA data. Hon'ble Vice Chancellor and Bioinformatics infrastructure facility (DBT funded), Sardar Vallabhbhai Patel University of Agriculture and Technology, Meerut- 250110, Uttar Pradesh for providing the facilities to complete research work.

References

1. Chakrabarti RD, Dinkins, Hunt AG. De novo Transcriptome Assembly and Dynamic Spatial Gene Expression Analysis in Red Clover. Plant Genome. 2015; 10:3835.
2. Chinnappa KS, McCurdy DW. *De novo* assembly of a genome-wide transcriptome map of *Vicia faba* (L.) for transfer cell research. Front. Plant Sci. 2015; 6:217.
3. CLC bio. CLC genomics workbench. Release 8.0. CLC bio, Aarhus, Denmark, 2015. http://resources.qiagenbioinformatics.com/manuals/clcmainworkbench/current/index.php?manual=Sequencing_data_analyses_Assembly.html.
4. Cock PJA, Fields CJ, Goto N, Michael L, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucl. Ac. Res. 2010; 38(6):1767-1771.
5. Fu N, Wang Q, Shen HL. De Novo Assembly, Gene Annotation and Marker Development Using Illumina Paired-End Transcriptome Sequences in Celery (*Apium graveolens* L.). PLoS ONE. 2013; 8(2).
6. Goyal M, Singh J, Kumar P, Sirohi A. Mechanistic insights into longan (*Dimocarpous longan Lour.*) transcriptome for physiological characterization for defensive genes and differential gene expression analysis with longan embryogenic callus transcriptome. Pl. Omics Jour. 2017; 10(05):219-231.
7. Hetalkumar JP. *De novo* RNA seq assembly and annotation of important legume-*Vicia sativa* L. International Conference on Transcriptomics, 2015.
8. Kotwal S, Kaul S, Sharma P, Gupta M, Shankar R, Jain M *et al*. De Novo Transcriptome Analysis of Medicinally Important Plantago ovata Using RNA-Seq. PLoS ONE. 2016; 11(3).
9. Liu B, Zhang N, Wen Y, Jin X, Yang J, Si H *et al*. Transcriptomic changes during tuber dormancy release process revealed by RNA sequencing in potato. J Biotechnol. 2015; 6996:1-14.
10. Lakshmanan M, Lim S, Mohanty B, Kim JK, Ha S, Lee D. Unravelling the light-specific metabolic and regulatory signatures of rice through combined in silico modelling and multi-omics analysis. Pl. Physiol. 2015; 169:3002-3020.
11. Ohta T, Nakazato T, Bono H. Calculating the quality of public high-throughput sequencing data to obtain a suitable subset for reanalysis from the Sequence Read Archive. Giga Science. 2017; 6:1-8.
12. Roy NS, Kim J, Choi AY, Ban YW, Park N, Park KC *et al*. Gen. Inform. 2018; 16(4).
13. Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R *et al*. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Gen. Res. 2010; 20:646-654.