

## **Whole genome assembly of the Little Millet (*Panicum sumatrense*) genome: A climate resilient crop species**

**Rasmita Rani Das<sup>1,2</sup>, Ashish Kumar<sup>3</sup>, Dushyant Singh Baghel<sup>3</sup>, Ajay K Mahato<sup>4</sup>, Nrisingha Dey<sup>1</sup>, Pulok K Mukherjee<sup>5</sup> and Seema Pradhan<sup>1\*</sup>**

<sup>1</sup>Institute of Life Sciences, NALCO Square, Chandrasekharapur, Bhubaneswar-751023 (India)

<sup>2</sup>Regional Centre for Biotechnology, NCR Biotech Science Cluster, Faridabad-121001 (India)

<sup>3</sup>Nucleome Informatics Pvt Ltd, HITEC City, Hyderabad, Telangana-500081(India)

<sup>4</sup>Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad, Telangana-500039 (India)

<sup>5</sup>Institute of Bioresources and Sustainable Development (IBSD)Takyelpat, Imphal, Manipur- 795001 (India)

### **\*Correspondence:**

Seema Pradhan

[seema@ils.res.in](mailto:seema@ils.res.in)

## Abstract

Little millet (*Panicum sumatrense*) belongs to one of the largest genera, *Panicum* comprising of 500 species that are distributed especially in tropical and sub-tropical Asia as well as Africa. This is a climate resilient crop that has the ability to adapt to adverse growing conditions, especially drought. Added to this is the better nutritional profile of little millet, which has higher iron and fibre contents as compared with rice. These characteristics make it an important crop species that would be pivotal for ensuring food security. Generating genomic resources for the species is significant and will have implications in crop improvement. Therefore, in the present study we report the *de novo* whole genome sequence of *P. sumatrense*. Long read sequencing and Hi-C based scaffolding resulted in a total of 279 scaffolds with N50 of 7.8Mb. The genome was annotated to predict protein coding genes and orthologous groups were analysed after comparison with various plant genomes. This high-quality genome assembly can be a valuable resource as future reference for genomic studies in this crop and related crop species.

**Keywords:** *Panicum sumatrense*; little millet; whole genome assembly; genome annotation; Simple sequence repeats

## Background

*Panicum sumatrense* (Little Millet) is a small seeded cereal crop mostly grown in the semi-arid tropical and sub-tropical regions of Asia and Africa. The genus *Panicum* L., comprising of about 500 species are distributed worldwide in tropical and subtropical regions, and is one of the largest genera of the family Poaceae [1]. *P. sumatrense* is a tetraploid ( $2n = 4x = 36$ ) [2] crop species and has some additional benefits due to their unique ability to grow in unfavorable ecological conditions like various abiotic stresses including adverse environments such as soil with high salinity and high temperatures [3]. Along with its ability to grow under unfavorable conditions, the species is highly nutritious with higher content of nutrients such as Iron and Fiber as compared to other staple crops like Rice and Wheat [4,5].

This nutri-cereal is a good source of food and feed across India and sub-Saharan Africa. The crop is consumed as a staple food in drought prone Indian states like Rajasthan, parts of Odisha, and is popularly known as “Indian Millet” [6]. Around 5,000 years ago, little millet was domesticated in India [7] and is grown mainly in certain Asian countries like India, Myanmar, Nepal and Sri Lanka [8]. The Indian little millet has a short crop cycle and can efficiently grow in both dry and waterlogged conditions with adverse environments such as soil with high salinity and temperatures. Its consistency in providing better yield in drought prone marginal lands enhances its importance as a crop for regional food stability and nutritional security. This versatile crop with better nutritional contents and health benefits with similar compositions as other cereal crops can be a worthy addition to the diet. Despite its nutritional, economical values, the crop is limited with genomic resources. In an attempt to generate some genetic database of little millet, our previous study tried to explore the transcriptome profile with the help of next-generation sequencing [6]. Although the chloroplast genome of little millet has earlier been reported [9], the whole genome remains to be sequenced yet.

We have previously reported the de novo transcriptome assembly of *P. sumatrense* cv. OLM 20 [6], and now have selected the same variety for whole genome sequencing and assembly with the help of third generation sequencing technology. With advent of high-throughput sequencing technology, the third-generation sequencing platforms have become new tools of genome research. These contemporary sequencing platforms circumvent some limitations posed by the second-generation technology such as reads being too short and may match with many different regions of the genome and are not unique to any specific region of the sequence. The significantly longer reads from platforms like the PacBio sequel II, have made genome assemblies comparatively error free and

especially effective for handling polyploid genomes [10-11]. In this study, we present the first report of the whole genome assembly of *P. sumatrense*.

## Methods

### Sampling and Sequencing

The seeds from a single plant of *P. sumatrense* OLM 20 were obtained from CPR, Berhampur, India (19° 18' 53.8632" N, 84° 47' 38.7240" E). They were grown in soil (local sandy loam soil) in controlled conditions (light/dark:16hr/8hr; 55% RH, 28°C  $\pm$ 2 temperature). Young leaves were collected from the plants at 5 leaf stage (Supplementary figure S1), surface sterilized with 70% ethanol and stored at -80°C after freezing in liquid N<sub>2</sub>. High molecular weight genomic DNA was isolated with QIAGEN 100G kit by following manufacturers protocol. DNA purity, size and concentration was recorded by using NanoDrop 2000 (Thermo Scientific), FEMTO Pulse (Agilent) and Qubit 3 fluorometer (Invitrogen). Good quality, high molecular weight genomic DNA was used for library preparation and sequencing. High molecular weight genomic DNA was sheared mechanically on Megaruptor 3 (Diagenode) to generate 30 Kb fragments. These fragments were used to prepare the library using SMRTBell Express template Preparation Kit2.0 (Pacific Biosciences) following manufacturer's protocol. The libraries were purified using AMPure PB beads (PACBIO) and size selected using BluePippin (Sage science). The quality filtered libraries were sequenced on PacBio Sequel II platform to generate a total of 317 GB of unique molecular yield data. Then the high molecular weight DNA was purified and used for library preparation for generating 1.01 GB of assembly HiC data using Proximity Ligation technology workflow (Phase Genomics)

### *De novo* genome assembly

Initially, the PacBio reads were screened for generation of Hifi reads using CCS. A primary genome assembly was generated using hifiasm with the option to integrate Hi-C reads (v0.16.1 r375) [12]. HiFi reads were aligned to the assembled genome using minimap2 (v2.24 r1122) [13] to create a coverage histogram. This coverage histogram was used to purge haplotigs from the assembly using purge\_haplotigs (v1.1.2) [14] and the contigs were collapsed based on this purging. HiC reads were aligned to this contig level assembly using BWA (v 0.7.17 r1188) [15]. Each pair was aligned separately to the assembly. The alignments were processed by removing chimeric alignments and duplicates followed by a MAPQ filter of 30 to retain only the best alignments. Scaffolding of the contigs was done using the SALSA pipeline (v2.3) which performs large scaffold construction using graph construction and link scoring function [16]. Misjoin, detection and correction was performed repeatedly until the best final scaffolds were achieved. Contig and scaffold level assemblies were assessed for completeness using BUSCO (v5.3.1) [17] and

assembly metrics were calculated using QUAST (v5.2.0) [18]. The final assembly (after scaffolding with Hi-C data) consisted of scaffolds with an average length of 2.03Mb (Table 1). The N50 value for assembled contigs/scaffolds was calculated to be 7.85 Mb, which is better than those reported for other millets species (Table 2).

**Table 1:** Assembly statistics after primary assembly with Hifiasm and improved assembly after Hi-C sequencing

| Assembly parameter              | PacBio reads assembled with Hi-C integrated Hifiasm | After purging | After scaffolding with Hi-C data |
|---------------------------------|---|---------------|----------------------------------|
| Number of Contigs/scaffolds     | 14282   | 433           | 279                              |
| Maximum contig/scaffold length  | 19976206 bp   | 19976206 bp   | 25599187 bp                      |
| Minimum contig/scaffold length  | 13971 bp  | 22837 bp      | 16609 bp                         |
| Average contig/scaffold size    | 144236.86 bp  | 1308524.51 bp | 2031115.10 bp                    |
| N50                             | 487704 bp   | 3677365 bp    | 7858070 bp                       |
| N50 index                       | 674   | 37            | 19                               |
| Total number of bases assembled | 2059990919  | 566591115     | 566681115                        |

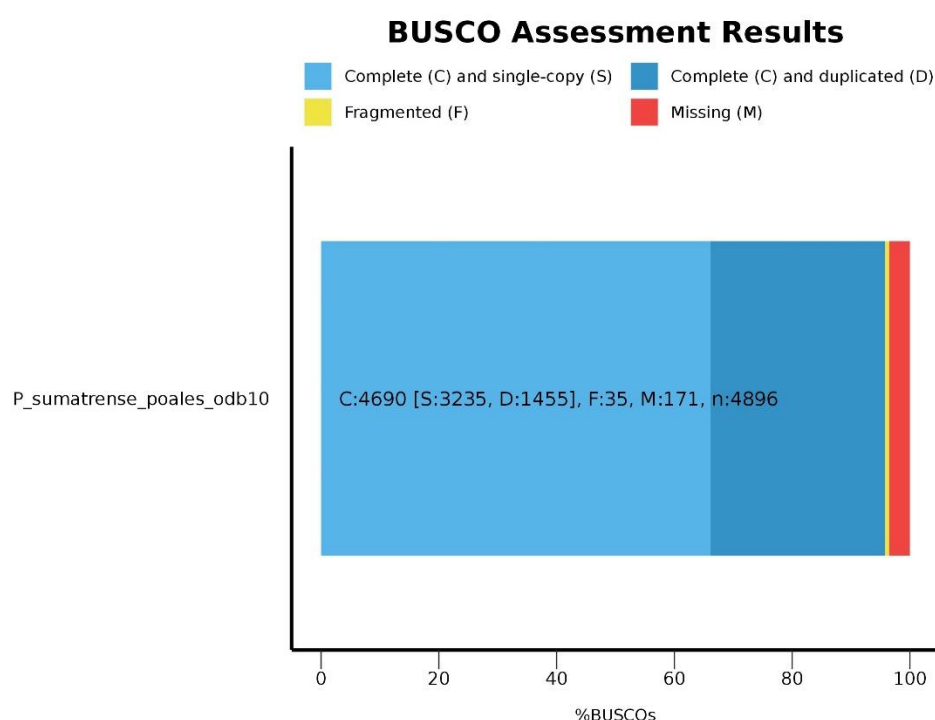
**Table 2:** Comparison with other millet genomes

|                     | Little Millet              | Finger Millet<br>(Hittalmani <i>et al</i> 2017) [19] | Foxtail Millet<br>(Zhang <i>et al</i> 2012) [20] | Pearl Millet<br>(Varshney <i>et al</i> 2017) [21] |
|---------------------|----------------------------|--|--|---|
| <b>Ploidy level</b> | Allotetraploid<br>2n=4x=36 | Allotetraploid<br>2n=4x=36                           | Diploid<br>2n=2x=18                              | Diploid<br>2n=2x=14                               |

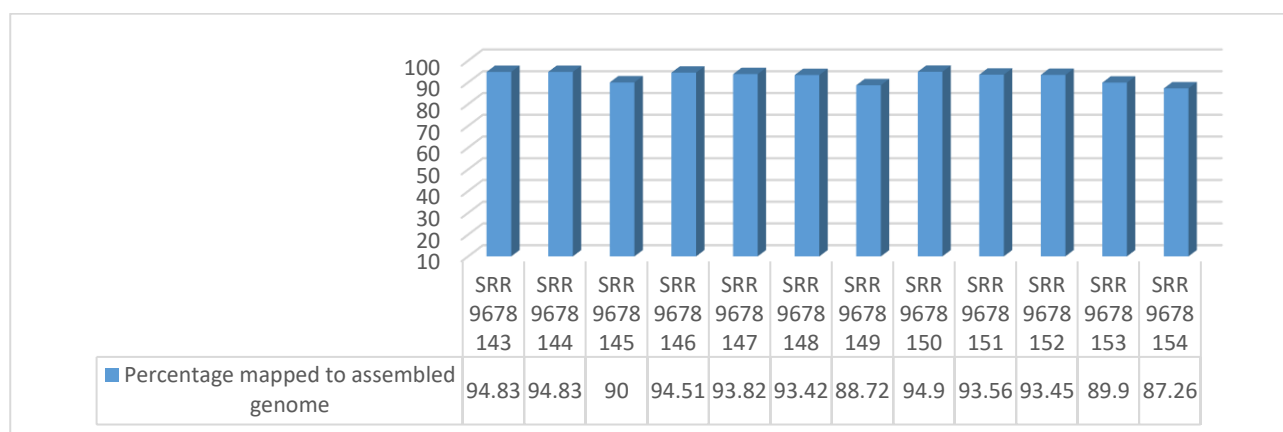
|   |         |           |       |           |
|---|---------|-----------|-------|-----------|
| <b>Estimated genome size (Gb)</b>                   | 1.2     | 1.59      | 0.493 | 1.76      |
| <b>Total length of the assembled sequences (Mb)</b> | 566.68  | 1196.06   | 423   | 1556.1    |
| <b>No. of contigs/scaffolds</b>                     | 279     | 525,759   | 136   | 175,708   |
| <b>N50</b>  | 7.85 Mb | 23,732 bp | 1Mb   | 18,180 bp |

## Genome quality assessment

The final assembly generated after scaffolding with HiC was assessed for quality using BUSCO (v5.3.1) [17]. We found a total of 4690 complete BUSCOs out of a total of 4896 BUSCOs from the poales\_odb10 database to be represented in this assembly of little millet genome (Figure 1). We also mapped the reads from RNA-Seq of little millet onto the assembled genome and found that, on an average, more than 90% of the reads aligned to the genome (Figure 2). The LTRs identified through LTR\_harvest [19] were analysed with LTR\_retriever [https://github.com/oushujun/LTR\_retriever] to obtain the LAI (LTR Assembly Index). The LAI for the whole genome assembly of *P. sumatrense* was found to be 18.26 (Supplementary data). The data from all the quality assessment methods/tools suggests that this assembly of the little millet genome is of a superior quality.



**Figure 1. BUSCO assessment results for little millet genome assembly using Poales\_odb10 database (C:95.8%[S:66.1%, D:29.7%],F:0.7%,M:3.5%,n:4896)**



**Figure 2. Percentage of reads from RNA-Seq data of Little millet in NCBI mapped onto the assembled genome.**

## Repeat elements and Genome annotation

RepeatModeler was also used to generate a customised repeat database for little millet. The resulting customized repeat database consisting of a total of 8950 sequences was used to screen the assembled genome using RepeatMasker program (v4.1.1, <http://www.repeatmasker.org/>) that screens DNA sequences for retroelements, interspersed repeats, transposons and low complexity DNA sequences (Table 3). Most of the repeat elements were “retroelements” (30.05%) and LTR elements were found to be the most abundant type of retroelements (29.05%). The GC content was found to be 45.86% and about 54% bases were masked. The protein sequences reported in *Panicum hallii* ([https://plants.ensembl.org/Panicum\\_hallii/Info/Index](https://plants.ensembl.org/Panicum_hallii/Info/Index)) were used to predict the protein coding sequences in *P. sumatrense* using BRAKER (<https://github.com/Gaius-Augustus/BRAKER>; [20]). This led to the prediction of 49,690 protein coding sequences in the genome of little millet, of which 43,961 sequences (88.4%) could be annotated with the Uniprot SwissProt database.

**Table 3: Statistics of genome repeat masking**

|                                     |                           |                        |                               |
|-------------------------------------|---------------------------|------------------------|-------------------------------|
| Sequences: 279                      |                           |                        |                               |
| Total length: 566681115 bp          |                           |                        |                               |
| GC level: 45.86%                    |                           |                        |                               |
| Bases masked: 308857442 bp (54.50%) |                           |                        |                               |
|                                     | <b>Number of elements</b> | <b>Length occupied</b> | <b>Percentage of sequence</b> |
| <b>Retroelements</b>                | <b>85473</b>              | <b>170312443 bp</b>    | <b>30.05%</b>                 |

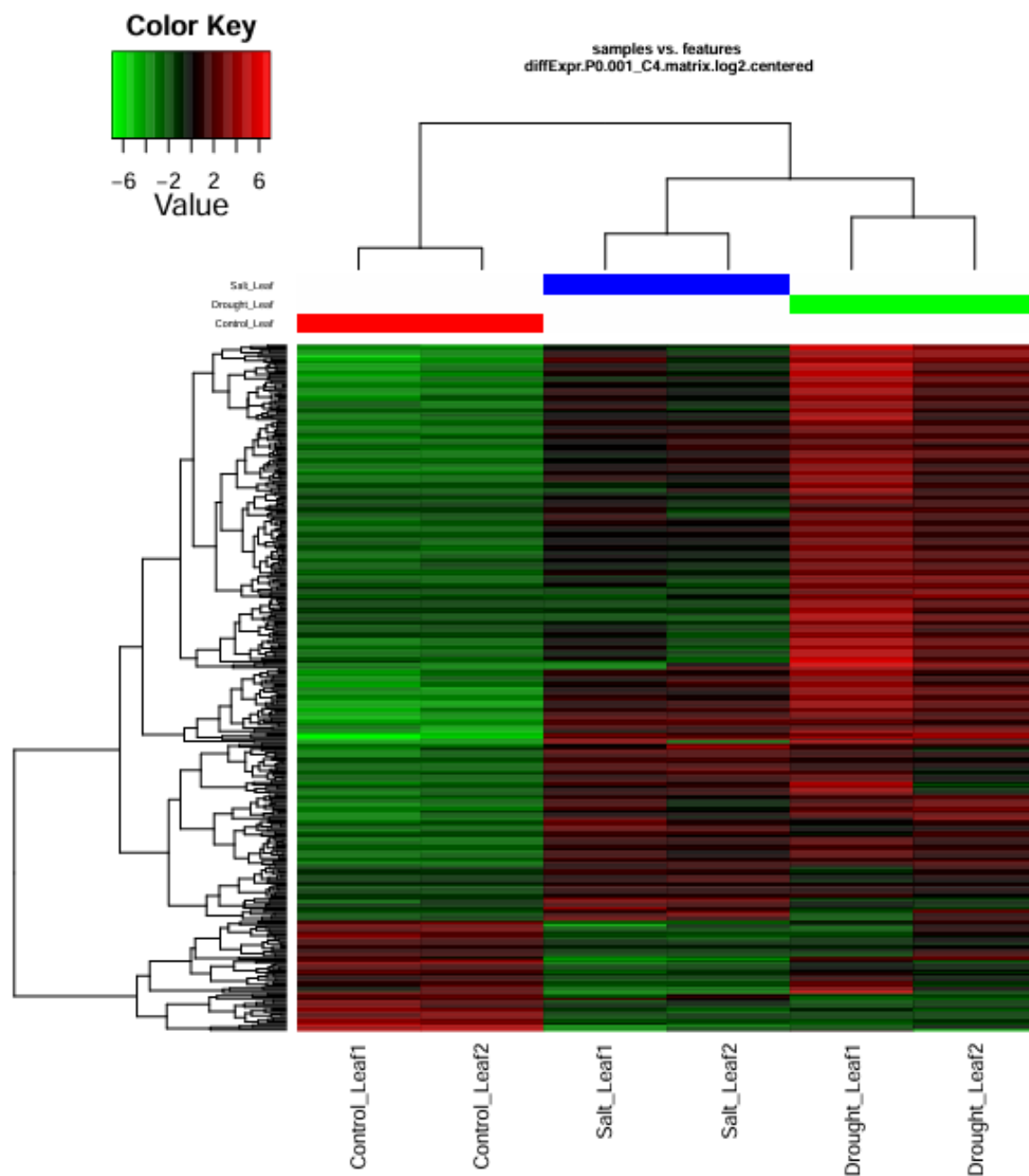
|                                   |               |                     |               |
|-----------------------------------|---------------|---------------------|---------------|
| SINEs                             | 0             | 0 bp                | 0.00%         |
| Penelope                          | 0             | 0 bp                | 0.00%         |
| LINEs                             | 11734         | 5670933 bp          | 1.00%         |
| CRE/SLACS                         | 0             | 0 bp                | 0.00%         |
| L2/CR1/Rex                        | 0             | 0 bp                | 0.00%         |
| R1/LOA/Jockey                     | 0             | 0 bp                | 0.00%         |
| R2/R4/NeSL                        | 0             | 0 bp                | 0.00%         |
| RTE/Bov-B                         | 757           | 259125 bp           | 0.05%         |
| L1/CIN4                           | 10977         | 5411808 bp          | 0.96%         |
| LTR elements                      | 73739         | 164641510 bp        | 29.05%        |
| BEL/Pao                           | 0             | 0 bp                | 0.00%         |
| Ty1/Copia                         | 16449         | 28221702 bp         | 4.98%         |
| Gypsy/DIRS1                       | 57222         | 136364159 bp        | 24.06%        |
| Retroviral                        | 0             | 0 bp                | 0.00%         |
| <b>DNA transposons</b>            | <b>16273</b>  | <b>10827300 bp</b>  | <b>1.91%</b>  |
| hobo-Activator                    | 2712          | 1540593 bp          | 0.27%         |
| Tc1-IS630-Pogo                    | 302           | 211507 bp           | 0.04%         |
| En-Spm                            | 0             | 0 bp                | 0.00%         |
| MuDR-IS905                        | 0             | 0 bp                | 0.00%         |
| PiggyBac                          | 0             | 0 bp                | 0.00%         |
| Tourist/Harbinger                 | 1734          | 704105 bp           | 0.12%         |
| Other                             | 0             | 0 bp                | 0.00%         |
| <b>Rolling-circles</b>            | <b>1600</b>   | <b>903838 bp</b>    | <b>0.16%</b>  |
| <b>Unclassified</b>               | <b>293564</b> | <b>121897832 bp</b> | <b>21.51%</b> |
| <b>Total interspersed repeats</b> | <b>--</b>     | <b>303037575 bp</b> | <b>53.48%</b> |
| <b>Small RNA</b>                  | <b>0</b>      | <b>0 bp</b>         | <b>0.00%</b>  |
| <b>Satellites</b>                 | <b>0</b>      | <b>0 bp</b>         | <b>0.00%</b>  |
| <b>Simple repeats</b>             | <b>93833</b>  | <b>4403677 bp</b>   | <b>0.78%</b>  |
| <b>Low complexity</b>             | <b>10280</b>  | <b>512352 bp</b>    | <b>0.09%</b>  |



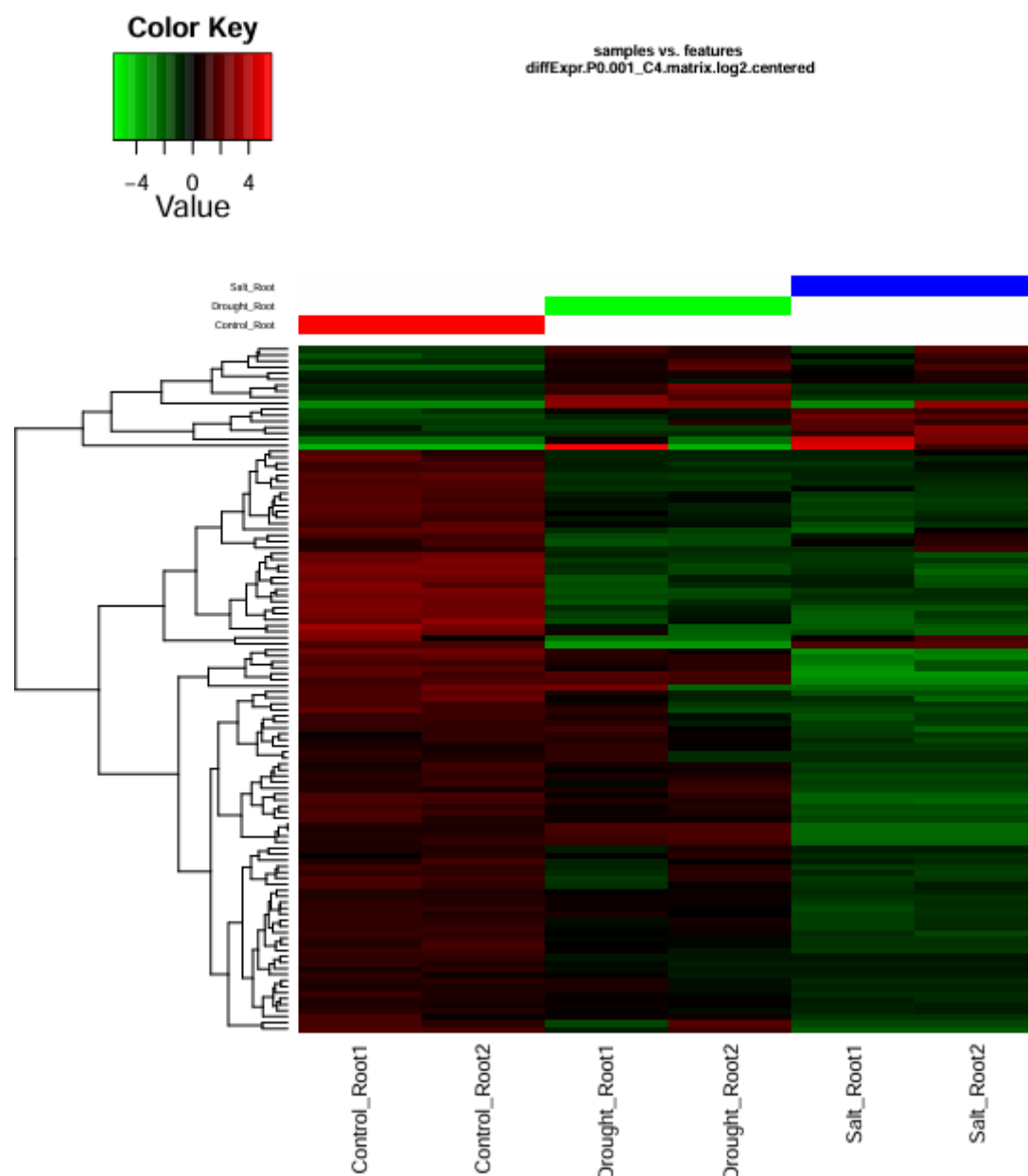
## Gene expression under drought and salinity stress

Analysing the orthologous groups present in plants that may be closely or distantly related provide insights into the probable evolution of a plant. The protein sequences predicted from the whole genome of *Panicum sumatrense* were compared with those from 8 other plants; *Oryza sativa (indica)*, *Brachypodium distachyon*, *Eleusine coracana*, *Panicum hallii*, *Panicum miliaceum*, *Panicum virgatum*, *Pennisetum glaucum*, and *Setaria italica* using OrthoFinder [21] to identify the orthologous groups shared by and specific to these plants. The analysis showed that a total of 442873 genes were assigned to 55,120 orthogroups (Supplementary data). Of these, 1834 orthogroups were specific to *P. sumatrense* (Supplementary data). The genes present in these little millet specific orthologous groups were grouped into various GO terms to get an idea about what distinguishes little millet from its relatives in the poaceae family. It was observed that processes like hydrolase (carbon-nitrogen bonds) activity, glucosidase activity and oxidoreductase activity under “GO Slim-Molecular function” category and processes related to metabolism of phosphate containing compounds and organonitrogen compounds under the “GO Slim-Biological processes” category were enriched in these orthologous groups.

Further, a gene expression atlas was generated after mapping the RNA-seq reads from a previous study [6] onto the complete set of protein coding genes predicted from the whole genome of little millet to identify differentially expressed genes in response to drought and salinity stresses. Briefly, the clean reads from each sample were mapped onto the genes using Bowtie2 [22] and it was observed that genes encoding LEA proteins (g4028.t1), Dehydrin (g12235.t1), seed maturation protein (g14378.t1), F-box proteins (g43661.t1, g6953.t1), E3 ubiquitin ligases (g11981.t2), pyruvate dehydrogenase (g39238.t1) and transcription factors encoding MYB (g31522.t1) and MADS (g29003.t1) TFs were upregulated in leaf tissue of little millet under conditions of drought and salinity stress (Figure 3A, Supplementary data). A similar analysis in root tissue led to an interesting observation where most of the genes were found to be downregulated in response to drought and salinity stress as compared with the control sample. This is in stark contrast to the results obtained in case of leaf tissue and showed a marked downregulation of genes encoding proteins like Early nodulin (g38606.t1), L-lactate dehydrogenase (g3894.t1), Stearoyl-[acyl-carrier-protein]9-desaturase (g6357.t1), Alcohol dehydrogenase (g20376.t1), heat shock protein (g4349.t1). On the other hand, genes coding for proteins like F-box/kelch-repeat protein (g1251.t1), WRKY transcription factor (g42302.t2), Asparagine synthetase (g25249.t2) and UDP-glycosyltransferase (g19604.t1) were upregulated in root affected by salinity stress while that coding for a probable glycosyltransferase (g35695.t2) was upregulated in root in response to drought stress (Figure 3B, Supplementary data).



**Figure 3A. Gene expression profile for *P. sumatrense* leaf tissue in response to drought and salinity stress.**



**Figure 3B. Gene expression profile for *P. sumatrense* root tissue in response to drought and salinity stress.**

### SSRs in the little millet genome

Simple sequence repeats (SSRs) are one of the most valuable molecular markers with varied applications, especially in determining genetic diversity and plant breeding. So, MISA software (<https://webblast.ipk-gatersleben.de/misa/>) was used to mine SSRs in the little millet genome with the following parameters: Unit size / minimum number of repeats: (1/10) (2/6) (3/4) (4/4) (5/3) (6/3); Maximal number of bases interrupting 2 SSRs in a compound microsatellite:100. A total of 79,830 SSRs were identified in 270 sequences (Table 4). Majority of these were mono-nucleotide repeats (42.5%) and tri-nucleotide repeats (34.4%).

**Table 4:** SSRs identified in the genome sequence of *P. sumatrense*.

| Results of Microsatellite search               | Value     | Unit size | Number of SSRs |
|--|-----------|-----------|----------------|
| Total number of sequences examined             | 279       | 1         | 33942          |
| Total size of examined sequences (bp)          | 566681115 | 2         | 5656           |
| Total number of identified SSRs                | 79830     | 3         | 27481          |
| Number of SSR containing sequences             | 270       | 4         | 2425           |
| Number of sequences containing more than 1 SSR | 265       | 5         | 7217           |
| Number of SSRs present in compound formation   | 4690      | 6         | 3109           |

## Discussion

Millets have gained significant popularity, especially in the last few decades, owing to their superior nutritive value and climate resilience, making them a viable food source for future food security. These are also a potential resource for identification of molecular markers and genes that could further the process of crop improvement. However, these endeavours are limited due to a lack of genomic resources in millets. While there are reports on genomic resources of other minor millet species like finger millet [23], foxtail millet [24], pearl millet [25], such studies on little millet are very limited. The large and polyploid genome size of this plant poses certain difficulties in whole genome assembly. This study is an attempt to assemble the complex polyploid genome of *P. sumatrense* which generated a high-quality whole genome assembly for this tetraploid plant species. The good quality of the little millet assembly was evident after comparison with other millet counterparts and reiterated by more than 95% of BUSCOs being represented in this genome assembly and more than 90% of short reads from RNA-Seq being mapped onto the assembled genome [6]. The 49,690 predicted genes coding for viable protein products in the little millet genome could serve as a source of genic molecular markers, which in turn would be valuable for marker assisted selections.

*P. sumatrense* is a marginally cultivated cereal crop that can withstand conditions of drought and salinity [26]. Thus, studying the pattern of gene expression in this plant could help identify genes that may impart tolerance to these devastating stresses that lead to heavy losses in susceptible crops like rice. Similar analysis has been reported earlier for *P. sumatrense* in a study of the transcriptome of the plant in response to drought and salinity stress [6]. However, gene models that arise out of whole genome assembly, provide a more holistic dataset for generating gene expression atlas. The comparison of these protein sequences encoded in the little millet genome with those from other plants provided us with insights into the evolutionary process that could have contributed to the ability of this plant to withstand significant levels of abiotic stress. The observations pointed to the involvement of processes regulating oxidation-reduction and nitrogen and phosphate compound metabolism in the plant's response to abiotic stresses. The gene expression data also indicated the

role of well characterised genes such as LEA, F-box, dehydrin etc. in combating drought and salinity stress. The activity of these genes could be a critical factor contributing to plant growth in difficult terrain and provide a source of nutrition to marginalised population. This study will be a substantial contribution to the field of plant genomics and will serve as useful genomic resources which could contribute to future crop breeding and enhancement. However, this is a first draft of the whole genome assembly of *Panicum sumatrense* and we are curating it to generate an improved reference-grade genome assembly for this valuable crop plant.

## **Additional files**

**See Supplementary data**

## **Abbreviations**

Bp: base pairs; Mb: mega-base pairs; Gb: giga-base pairs; PacBio: Pacific Biosciences; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Aligner; LTR: long terminal repeat; NCBI: National Center for Biotechnology Information; BLAST: Basic Local Alignment Search Tool; SSR: Simple Sequence Repeat

## **Competing interests**

The authors declare that they don't have any competing interests.

## **Authors' contributions**

R.R.D collected the plant materials; R.R.D, S.P, D.S.B, A.K. designed the project; S.P., D.S.B, A.K and A.K.M worked on sequencing and data analysing; R.R.D, P.K.M, N.D and S.P wrote the manuscript; All the authors revised and approved the final version of the manuscript.

## **Acknowledgements**

All authors acknowledge the contributions of Late Dr. Ajay Parida, who initiated the study and inspired everyone for the execution of the study. RRD is thankful to Department of Science and Technology (DST-INSPIRE), India for her fellowship. The study was carried out with core funds provided by Department of Biotechnology, Ministry of Science and Technology, Government of India.

## References

1. Aliscioni SS, Giussani LM, Zuloaga FO, Kellogg EA. A molecular phylogeny of *Panicum* (Poaceae: Paniceae): tests of monophyly and phylogenetic placement within the Panicoideae. *American Journal of Botany*. 2003 May;90(5):796-821.
2. Saha D, Gowda MC, Arya L, Verma M, Bansal KC. Genetic and genomic resources of small millets. *Critical Reviews in Plant Sciences*. 2016 Jan 2;35(1):56-79.
3. Ajithkumar IP, Panneerselvam R. ROS scavenging system, osmotic maintenance, pigment and growth status of *Panicum sumatrense* roth. under drought stress. *Cell biochemistry and biophysics*. 2014 Apr;68(3):587-95.
4. Muthamilarasan M, Dhaka A, Yadav R, Prasad M. Exploration of millet models for developing nutrient rich graminaceous crops. *Plant Science*. 2016 Jan 1;242:89-97.
5. Shyamli PS, Rana S, Suranjika S, Muthamilarasan M, Parida A, Prasad M. Genetic determinants of micronutrient traits in graminaceous crops to combat hidden hunger. *Theoretical and Applied Genetics*. 2021 Jun 6:1-9.
6. Das RR, Pradhan S, Parida A. De-novo transcriptome analysis unveils differentially expressed genes regulating drought and salt stress response in *Panicum sumatrense*. *Scientific reports*. 2020 Dec 4;10(1):1-4.
7. de Wet JM, Brink DE, Rao KP, Mengesha MH. Diversity in kodo millet, *Paspalum scrobiculatum*. *Economic Botany*. 1983 Apr;37(2):159-63.
8. Prasada Rao KE, De Wet JM, Gopal Reddy V, Mengesha MH. Diversity in the small millets collection at ICRISAT. *Advances in small millets*. Edited by KW Riley, SC Gupta, A. Seetharam, and JN Mushonga. Oxford & IBH Publishing Co. Pvt. Ltd., New Delhi, India. 1993.
9. Sebastin R, Lee GA, Lee KJ, Shin MJ, Cho GT, Lee JR, Ma KH, Chung JW. The complete chloroplast genome sequences of little millet (*Panicum sumatrense* Roth ex Roem. and Schult.) (Poaceae). *Mitochondrial DNA Part B*. 2018 Jul 3;3(2):719-20.
10. Novikova PY, Tsuchimatsu T, Simon S, Nizhynska V, Voronin V, Burns R, Fedorenko OM, Holm S, Säll T, Prat E, Marande W. Genome sequencing reveals the origin of the allotetraploid *Arabidopsis suecica*. *Molecular biology and evolution*. 2017 Apr 1;34(4):957-68.
11. Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, Ding M, Ye W, Kirkbride RC, Jenkins J, Plott C. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. *Nature genetics*. 2020 May;52(5):525-33.
12. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*. 2021 Feb;18(2):170-5.
13. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018 Sep 15;34(18):3094-100.
14. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics*. 2018 Dec;19(1):1-0.
15. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*. 2013 Mar 16.
16. Ghurye J, Pop M, Koren S, Bickhart D, Chin CS. Scaffolding of long read assemblies using long range contact information. *BMC genomics*. 2017 Dec;18(1):1-1.

17. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015 Oct 1;31(19):3210-2.
18. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013 Apr 15;29(8):1072-5.
19. Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics*, 9, 1-14.
20. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics*, 3(1), lqaa108.
21. Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20, 1-14.
22. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), 1-10.
23. Hittalmani S, Mahesh HB, Shirke MD, Biradar H, Uday G, Aruna YR, Lohithaswa HC, Mohanrao A. Genome and transcriptome sequence of finger millet (*Eleusine coracana* (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. *BMC genomics*. 2017 Dec;18(1):1-6.
24. Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, Tao Y. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature biotechnology*. 2012 Jun;30(6):549-54.
25. Varshney RK, Shi C, Thudi M, Mariac C, Wallace J, Qi P, Zhang H, Zhao Y, Wang X, Rathore A, Srivastava RK. Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nature biotechnology*. 2017 Oct;35(10):969-76.
26. Wilson, M. L., & VanBuren, R. (2022). Leveraging millets for developing climate resilient agriculture. *Current Opinion in Biotechnology*, 75, 102683.