The draft genome of *Corchorus olitorius* cv. JRO-524 (Navin)

Debabrata Sarkarb,1, Ajay Kumar Mahatoa,1, Pratik Satya, Abjit Kundu, Sangeeta Singh, Pawan Kumar Jayaswal, Akshay Singh, Kaushlendra Bahadur, Sasmita Pattnaik, Nisha Singh, Avrajit Chakraborty, Nur Alam Mandal, Debajeet Das, Tista Basu, Amitha Mithra Sevanthia, Dipnarayan Saha, Subhojit Datta, Chandan Sourav Kar, Jiban Mitra, Karabi Datta, Pran Gobinda Karmakarb, Tilak Raj Sharmaa, Trilochan Mohapatrade, Nagendra Kumar Singha,⁎

a ICAR-National Research Centre on Plant Biotechnology (NRCPB), IARI, Pusa Campus, New Delhi 110012, India
b ICAR-Central Research Institute for Jute and Allied Fibres (CRIJAF), Nilganj, Barrackpore, Kolkata 700120, West Bengal, India
c Plant Molecular Biology and Biotechnology Laboratory, Department of Botany, University of Calcutta, Kolkata 700019, West Bengal, India
d Secretary (DARE) & Director General (ICAR), New Delhi 110001, India

**ARTICLE INFO**

**Keywords:**
Bast fibre  
Corchorus olitorius  
Dark jute  
Illumina MiSeq  
Whole genome sequence

**ABSTRACT**

Here, we present the draft genome (377.3 Mbp) of *Corchorus olitorious* cv. JRO-524 (Navin), which is a leading dark jute variety developed from a cross between African (cv. Sudan Green) and indigenous (cv. JRO-632) types. We predicted from the draft genome a total of 57,087 protein-coding genes with annotated functions. We identified a large number of 1765 disease resistance-like and defense response genes in the jute genome. The annotated genes showed the highest sequence similarities with that of *Theobroma cacao* followed by *Gossypium raimondii*. Seven chromosome-scale genetically anchored pseudomolecules were constructed with a total size of 8.53 Mbp and used for synteny analyses with the cocoa and cotton genomes. Like other plant species, gypsy and copia retrotransposons were the most abundant classes of repeat elements in jute. The raw data of our study are available in SRA database of NCBI with accession number SRX1506532. The genome sequence has been deposited at DDBJ/EMBL/GenBank under the accession LLWS00000000, and the version described in this paper will be the first version (LLWS01000000).

**Specifications**

| Organism/cell line/tissue | Dark jute (*Corchorus olitorius* cv. JRO-524)/ leaves |
|---------------------------|------------------------------------------------------|
| Sex                       | Hermaphrodite                                        |
| Sequence or array type    | Illumina MiSeq                                       |
| Data format               | Raw and processed                                    |
| Experimental factors      | The draft genome sequence of *Corchorus olitorius* cv. JRO-524 (Navin) |
| Experimental features     | DNA was extracted from seedling leaves of *C. olitorius* cv. JRO-524, and shotgun libraries were prepared followed by paired-end sequencing on an Illumina MiSeq platform, generating 2 × 250 bp overlapping reads. The cleaned sequence reads were merged with PANDASeq and assembled de novo using Newbler software. |

**Consent**

N/A

**Sample source location**

Barrackpore, Kolkata, India (22°46′2.7372″ N 88°23′18.0384″ E)

**1. Direct link to deposited data**

[http://www.ncbi.nlm.nih.gov/bioproject/PRJNA278717](http://www.ncbi.nlm.nih.gov/bioproject/PRJNA278717) for *Corchorus olitorius* cv. JRO-524 ([http://www.ncbi.nlm.nih.gov/sra/SRX1506532](http://www.ncbi.nlm.nih.gov/sra/SRX1506532)).

([https://www.ncbi.nlm.nih.gov/biosample/SAMN04160039](https://www.ncbi.nlm.nih.gov/biosample/SAMN04160039)).
2. Introduction

Corchorus olitorius L. (2n = 2 × 14; Malvaceae s. l.), commonly known as dark jute or jute mallow, is an important ligno-cellulosic bast fibre crop, with > 80% acreage of jute growing areas of the world. Grown in tropical lowland areas, it produces one of the strongest vegetable fibres and is only next to cotton in terms of production [1]. Though it is ideally suited for transplanted paddy-based crop rotation and makes softer and stronger fibre than its other cultivated counterpart C. capsularis (white jute), there are several biological constraints that limit its diversified uses in textile industry [2]. Besides yield enhancement, there is an urgent need to develop dark jute varieties with quality fibre in terms of fibre fineness and tensile strength including low-lignin content using genomics-assisted breeding approaches. Recently, the draft genome sequence of C. olitorius cv. O-4 has been released by Bangladesh [3]. However, the variety sequenced by Bangladesh is a pure line selection from a local landrace [4]. Since C. olitorius originated in Africa [5] and reached India together with many African crops in prehistory [6], it is of potential interest to decode one of its genomes that represents an admixture of both African and Indian gene pools. In this study, we sequenced a leading Indian variety JRO-524 (Navin), which was developed from a cross between African (cv. Sudan Green from Sudan) and indigenous (cv. JRO-632; a local selection) types. Our results provide new insights into the C. olitorius genome, and its availability would not only facilitate jute research and development, but also foster the application of translational genomics in jute improvement.

3. Experimental design, material and methods

3.1. Plant material and DNA isolation

Seeds of C. olitorius cv. JRO-524 were germinated in petri dishes and leaves were collected from 10-day-old seedlings. Twenty leaves collected from ten seedlings were pooled and used for DNA extraction using the GenElute™ Plant Genomic DNA Miniprep Kit (Sigma-Aldrich Co., St. Louis, USA).

3.2. Genome sequencing, de novo assembly and annotation

DNA was fragmented using the Covaris AFA™ system (Covaris, Inc., Woburn, USA) with a median fragment size of 544 bp, and shotgun libraries were prepared using the Illumina TruSeq DNA PCR-Free Sample Preparation Kit (Illumina, San Diego, USA). Paired-end sequencing was performed on two flow cells of an Illumina MiSeq (2 × 250 bp) platform. The sequence reads were quality-checked using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Following adapter trimming, the poor-quality bases were removed using Trimmomatic v.0.36 [7]. The genome size was evaluated using the K-mer Analysis Toolkit (KAT) [8]. High-quality reads were merged using PANDASeq v2.7 [9], and then assembled de novo using Newbler v. 2.6 with default parameters (Roche Inc. Germany). We used the FGENESH gene prediction pipeline from the software package Molquest v4.5 (http://www.softberry.com) for the in silico prediction of genes. The predicted genes were annotated using BLASTX (E < 10⁻5) search against the NCBI non-redundant (nr) protein database.

3.3. Synteny mapping and pseudomolecule construction

SyMap v3.4 [10] was used for pairwise synteny mapping with cocoa (Theobroma cacao) and diploid cotton (Gossypium raimondii) that showed the highest sequence similarities with our assembled C. olitorius genome during the BLAST similarity search. For the construction of seven chromosome-scale pseudomolecules, we used ALLMAPS [11] to integrate the genome assembly with a RAD-SNP-based genetic map of C. olitorius [12].

3.4. Identification of disease resistance-like and defense response genes

The disease resistance-like (R-like) and defense response (DR) genes were manually categorized using different keywords/phrases that represent R-like and DR genes into five main classes as follows: (i) NBS-LRR (matching with NBS-LRR and LRR, CC-NBS-LRR, Pib, Pita, Rp 1-d8, Lr10, Mla 1 and rust resistance), (ii) LZ-NBS-LRR (matching with LZ-NBS-LRR, but not with NBS-LRR, CC-NBS-LRR, LRR and RPM1), (iii) LRR-TM (matching with Xa21, serine/threonine kinases and C2/C5S resistance), (iv) LRR (matching with disease resistance, viral resistance, Yr10, LRR, but not with NBS-LRR, CC-NBS-LRR, LZ-NBS-LRR, and (v) defense response genes (matching with glucanases, chitinases and thaumatin like genes) [13]. We mapped these R-like and DR genes to an integrated RAD-SNP-based genetic map of jute [12].

3.5. Repeat elements and SSR identification

All assembled contigs were screened for the presence of simple sequence repeats (SSRs) using MISA (http://pgrc.ipk-gatersleben.de/ misa/). The assembled contigs were analyzed to identify repeat sequences using RepeatModeler and RepeatMasker with Repbase library v22.01 [14].

4. Data description

Illumina MiSeq sequencing generated 52,507,986 overlapping 2 × 250 bp paired-end raw reads (~15.65 Gbp sequence) that were processed to yield 24,996,514 merged high-quality reads with an average read length of 450 bp (~12.9 Gbp) and a 31.32× coverage of the K-mer based estimated 415 Mbp genome of C. olitorius cv. JRO-524. The longer merged reads from Illumina MiSeq platform facilitated economical de-novo assembly of jute genome into 52,373 contigs (377.3 Mbp) covering 90.8% of the estimated genome size. The mean contig size was 7206 bp, while the N50 size was 16,573 bp (Table 1). The raw sequence data are available in NCBI SRA database with accession number SRX1506532, and the assembled genome sequence has been deposited at DDBJ/EMBL/GenBank with the accession number LLWS00000000 vide BioProject PRJNA278717 and BioSample SAMN04160639. We predicted 76,881 gene models, with an average and the largest gene size of 1.3 kbp and 37 kbp, respectively. In total 59,531 (77.4%) of the predicted genes were annotated using BLASTx, while 17,350 genes (22.6%) remained non-annotated and were thus unique to C. olitorius cv. JRO-524 genome. Of these, 57,087 were protein-coding genes with annotated functions. The predicted genes showed the highest sequence similarity with that of T. cacao (37.45%), followed by G. raimondii (9.68%). Using a restriction site-associated DNA (RAD)-SNP linkage map, we have shown earlier that C. olitorius has the maximum syntenic relationship with cocoa followed by diploid cotton [12]. Recently, Islam et al. [3] have also reported the same.

Table 1

| Index         | Statistics          |
|---------------|---------------------|
| Raw reads     | 52,507,986          |
| High-quality merged reads | 24,996,514        |
| Number of assembled contigs | 52,373            |
| Size of assembled contigs (bp) | 377,376,943     |
| Longest contig (bp) | 177,749           |
| Shortest contig (bp) | 500               |
| Number of contigs > 1 kb | 41,086            |
| Number of contigs > 10 kb | 11,958            |
| Number of contigs > 100 kb | 36               |
| Mean contig size (bp) | 7206              |
| Contig N50 (bp) | 16,573             |
Table 2
Summary of seven chromosome-scale pseudomolecules of *C. olitorius* cv. JRO-524. The assembled genome was integrated with a RAD-SNP-based genetic map of *C. olitorius* [12] and anchored contigs were joined together with 50 Ns to generate the chromosome-scale pattern of syntenic relationship for assembled genome was integrated with a RAD-SNP-based genetic map of *C. olitorius* cv. JRO-524. The seven genetically anchored jute chromosomes with 10 chromosomes (8.53 Mbp) of genetically anchored jute genome revealed chromosomal level synteny of jute with both cocoa and cotton genomes.

| Chromosome | No. of RAD-SNP markers in genetic map | No. of mapped RAD-SNP markers in genome | No. of anchored contigs | Size of anchored contigs (bp) |
|------------|---------------------------------------|----------------------------------------|-------------------------|-------------------------------|
| Chr1       | 139                                   | 139                                    | 76                      | 2,336,828                     |
| Chr2       | 119                                   | 119                                    | 65                      | 1,979,308                     |
| Chr3       | 114                                   | 114                                    | 69                      | 2,035,515                     |
| Chr4       | 48                                    | 47                                     | 38                      | 742,950                       |
| Chr5       | 32                                    | 32                                     | 17                      | 582,942                       |
| Chr6       | 29                                    | 29                                     | 6                       | 400,300                       |
| Chr7       | 22                                    | 21                                     | 17                      | 441,461                       |
| Total      | 503                                   | 501                                    | 288                     | 8,519,304                     |

5. Conclusions

To our knowledge, the work presented here is the first whole genome sequence for a *C. olitorius* genotype derived from an African jute. *C. olitorius* cv. Sudan Green, one of the parents of cv. JRO-524, was primarily used to transfer premature flowering resistance (in early sowing) to indigenous types [18]. Thus an in-depth comparison of the present sequence with the recently published draft genome [3], would provide new insights that could help understand the mechanisms underlying premature flowering vis-à-vis photoperiodic control of bast fibre development in jute. This would allow breeding of high-yielding varieties with durable premature flowering resistance, which has been recently observed to be breaking down when dark jute crops are sown early under long-day conditions, possibly due to climate change.

Conflict of interest

The authors declare that they have no conflict of interests.

Acknowledgments

We thank ICAR-NPTC (Indian Council of Agricultural Research-Network Project on Transgenics in Crops) for financial support (sub-
project grant ID - ICAR-NPTC-3070). We acknowledge NxGenBio Life Sciences, New Delhi for assistance in Illumina MiSeq sequencing and raw data processing.

References

[1] B.S. Mahapatra, S. Mitra, M. Kumar, A.K. Ghorai, S.K. Sarkar, C.S. Kar, D.K. Kundu, P.G. Karmakar, An overview of research and development in jute and allied fibre crops in India, Indian J. Agron. 57 (2012) 72-82.

[2] D. Sarkar, P. Satya, N.A. Mandal, D. Das, P.G. Karmakar, N.K. Singh, Jute genomics: emerging resources and tools for molecular breeding, in: K.G. Ramawat, M.R. Ahuja (Eds.), Fiber Plants - Biology, Biotechnology and Applications, Springer International Publishing AG, Cham, 2016, pp. 155-200.

[3] M. Islam, J.A. Saito, E. Emad, B. Ahmed, M. Islam, A. Halim, Q. Hosen, M. Hossain, R. Ahmed, M. Hossain, S. Kabir, M. Khan, M. Khan, R. Hasan, N. Akhtar, U. Honi, R. Islam, M. Rashid, X. Wan, S. Hou, T. Haque, M. Azam, M. Moosa, S.M. Elias, M.A.M. Hasan, N. Mahmood, M. Shafiuddin, S. Shahid, N. Shommu, S. Jahan, S. Roy, A. Chowdhury, A. Akhand, G. Nisho, K. Uddin, T. Rebeya, E.S.M. Hoque, A. Snigdha, S. Morteza, S. Maini, M. Islam, M.Z.H. Laskar, M. Zaman, A. Yuryev, M. Uddin, M. Rahman, M. Haque, M. Alam, H. Khan, M. Alam, Comparative genomics of two jute species and insight into fibre biogenesis, Nat. Plants 3 (2017) 16223.

[4] H.S. Balyan, B.S. Mahapatra, P.K. Gupta, Origins of white (Corchorus capsularis L.) jute: a reevaluation based on nuclear and chloroplast microsatellites, J. Plant Biochem. Biotechnol. 22 (2013) 372-381.

[5] W. Bao, K.K. Kojima, R. Begum, F. Zakrzewski, G. Menzel, B. Weber, S.S. Alam, T. Schmidt, Comparative molecular cytogenetic analyses of a major tandemly repeated DNA family and darkness with application to plant genomes, Nucleic Acids Res. 39 (2011) e68.

[6] C. Soderlund, M. Bomhoff, W.M. Nelson, SyMAP v3.4: a turnkey synteny system with application to plant genomes, BMC Bioinformatics 13 (2012) 31.

[7] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina MiSeq sequence data, Bioinformatics 30 (2014) 2114–2120.

[8] A.P. Masella, A.K. Bartram, J.M. Truszkowski, D.G. Brown, J.D. Neufeld, PANDAseq: paired-end assembler for illumina sequences, BMC Bioinformatics 13 (2012) 31.

[9] A. Kundu, A. Chakraborty, N.A. Mandal, D. Das, P.G. Karmakar, N.K. Singh, D. Sarkar, A restriction-site-associated DNA (RAD) linkage map, comparative genomics and identification of QTL for histological fibre content coincident with those for retted bast fibre yield and its major components in jute (Corchorus olitorius L., Malvaceae s. l.), Mol. Breed. 35 (2015) 19.

[10] H. Tang, X. Zhang, C. Miao, J. Zhang, R. Ming, J.C. Schnable, P.S. Schnable, E. Lyons, J. Lu, ALLMAPS: robust scaffold ordering based on multiple maps, Genome Biol. 16 (2015) 1.

[11] A. Kundu, A. Chakraborty, N.A. Mandal, D. Das, P.G. Karmakar, N.K. Singh, D. Sarkar, A restriction-site-associated DNA (RAD) linkage map, comparative genomics and identification of QTL for histological fibre content coincident with those for retted bast fibre yield and its major components in jute (Corchorus olitorius L., Malvaceae s. l.), Mol. Breed. 35 (2015) 19.

[12] A. Kundu, A. Chakraborty, N.A. Mandal, D. Das, P.G. Karmakar, N.K. Singh, D. Sarkar, A restriction-site-associated DNA (RAD) linkage map, comparative genomics and identification of QTL for histological fibre content coincident with those for retted bast fibre yield and its major components in jute (Corchorus olitorius L., Malvaceae s. l.), Mol. Breed. 35 (2015) 19.

[13] S. Singh, S. Chand, N.K. Singh, T.R. Sharma, Genome-wide distribution, organisation and functional characterization of disease resistance and defence response genes across rice species, PloS One 10 (2015) e0125964.

[14] W. Bao, K.K. Kojima, O. Kohany, Repbase update, a database of repetitive elements in eukaryotic genomes, Mob. DNA 6 (2015) 11.

[15] X. Argout, J. Salse, J.-M. Aury, M.J. Guiltnan, G. Droz, J. Gouy, M. Allegre, C. Chaparro, T. Legavre, S.N. Maximova, M. Abrouk, F. Murat, O. Fouet, J. Pouliain, M. Ruiz, Y. Roquet, M. Rodier-Gourd, J.F. Barbosa-Neto, F. Sabot, D. Kudrna, J.S.S. Ammiraju, S.C. Schuster, J.E. Carlson, E. Sallet, T. Schiex, A. Dievet, M. Kramer, L. Gelley, Z. Shi, A. Berard, C. Viot, M. Boccara, A.M. Risterucci, V. Guignon, X. Sahau, M.J. Axtell, Z. Ma, Y. Zhang, S. Brown, M. Bourge, W. Golser, X. Song, D. Clement, R. Rivallan, M. Tah, J.M. Akaza, B. Pitollat, K. Gramacho, A. D'Hont, D. Brunel, D. Infante, I. Kebe, P. Costet, R. Wing, W.R. McCombie, E. Guidorodni, F. Questier, O. Panand, P. Wincker, S. Bocs, C. Lanaud, The genome of Theobroma cacao, Nat. Genet. 43 (2011) 101–108.

[16] K. Wang, Z. Wang, F. Li, W. Ye, J. Wang, G. Song, Z. Yue, L. Cong, H. Shang, S. Zhu, C. Zou, Q. Li, Y. Yuan, C. Lu, H. Wei, C. Gou, Z. Zheng, Y. Yin, X. Zhang, K. Liu, B. Wang, C. Song, N. Shi, R.J. Kohel, R.G. Percy, J.Z. Yu, Y.-X. Zhu, J. Wang, S. Yu, The draft genome of a diploid cotton Gossypium raimondii, Nat. Genet. 44 (2012) 1103–1108.

[17] R. Begum, P. Zakrzewski, G. Menzel, B. Weber, S.S. Alam, T. Schmidt, Comparative molecular cytogenetic analyses of a major tandemly repeated DNA family and retrotransposon sequences in cultivated jute Corchorus species (Malvaceae), Ann. Bot. 112 (2013) 123–134.

[18] C.S. Kar, P. Satya, J. Mitra, D. Sarkar, M.K. Sinha, A. Kundu, B.S. Mahapatra, Varietal development of jute and allied fibres in India, Indian Farm 60 (2010) 5–9.