Development of an InDel polymorphism database for jute via comparative transcriptome analysis

Journal:	<i>Genome</i>
Manuscript ID	gen-2017-0191.R1
Manuscript Type:	Article
Date Submitted by the Author:	24-Jan-2018
Complete List of Authors:	Yang, Zemao; Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences Dai, Zhigang; Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences Xie, Dongwei; Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences Chen, Jiquan; Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences Tang, Qing; Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences Cheng, Chaohua; Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences Xu, Ying; Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences Wang, Tingzhang; Zhejiang Institute of Microbiology Su, Jianguang; Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences, germplasm lab
Is the invited manuscript for consideration in a Special Issue? :	N/A
Keyword:	Marker, high-throughput sequencing, InDel, Corchorus

SCHOLARONE™
Manuscripts

Development of an InDel polymorphism database for jute via comparative transcriptome analysis

Zemao Yang¹, Zhigang Dai¹, Dongwei Xie¹, Jiquan Chen¹, Qing Tang¹, Chaohua Cheng¹, Ying Xu¹, Tingzhang Wang², Jianguang Su^{1*}

1, Institute of Bast Fiber Crops, Chinese Academy of Agricultural Sciences/Key laboratory of Stem-fiber Biomass and Engineering Microbiology, Ministry of Agriculture, Changsha, China

2, Zhejiang Institute of Microbiology, Hangzhou, 310012, China.

*Corresponding author

E-mail: zhongzhiziyuan@aliyun.com; jgsu@vip.163.com

telephone and fax numbers:0086-0731-88998519

Abstract: Jute (*Corchorus* spp.) is one of the most commercially important bast fiber crops in the world. However, molecular markers and high-density genetic maps are still lacking on jute compared with other crops. Insertion/deletion (InDel) markers, one of the most abundant sources of DNA/RNA variations in plant genomes, can easily be distinguished among different accessions using high-throughput sequencing. Using three transcriptome datasets, we identified and developed InDel markers. Altogether, 51 172 InDel sites in 18 800 unigenes were discovered, and the number of InDel loci per unigene varied from 1 to 31. Further, we found 94 InDel types, varying from 1 to 159 bp; the most common, were single-nucleotide (23 028), binucleotide (9824), and trinucleotide (9182). In total, 49 563 InDels in 18 445 transcripts were discovered in the comparison between TC and YG, followed by 48 934 InDels in 18 408 transcripts between NY and YG, and 3570 InDels in 2701 unigenes between NY and TC; Additionally, there were 1273 InDel sites in 1129 unigenes with polymorphisms between any two of the three accessions. Twenty-nine (58%) primer pairs represented polymorphisms when compared to the jute accessions, and PIC varied from 0.340 to 0.680, with an average of 0.491.

Keywords: Marker, high-throughput sequencing, InDel, *Corchorus*

Introduction

Jute (*Corchorus* spp.) is one of the most commercially important bast fiber crops in the world as it provides biodegradable and renewable lignocellulose fiber. The genus includes approximately 50–60 species distributed in the tropics, subtropics, and warm temperate regions of the world (mainly in Asia and Africa)(Ghosh et al. 2015). *Corchorus capsularis* L. (genome = 445 Mb) and *Corchorus olitorius* L. (genome = 338 Mb)(Islam et al. 2017), both self-pollinated diploid crops ($2n = 14$), are two most popular cultivated *Corchorus* species. With the development of high-throughput sequencing technology, the development of molecular biology in jute has reached a golden age. And some molecular markers, including random amplified polymorphic DNA (RAPD), sequence-related amplified polymorphisms (SRAPs), inter simple sequence repeats (ISSRs), simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs), have been developed and applied in jute for QTL mapping and genetic diversity and population genetics analyses(Mir et al. 2009; Ghosh et al. 2010; Mir et al. 2011; Banerjee et al. 2012;Topdar et al. 2013; Satya et al. 2016). However, the number of molecular markers are restricted in the most of researches , and to date, a few of molecular markers developed by high-throughput sequencing technology have been reported. Zhang et al. developed 1,906 EST-SSR markers using a transcriptome sequence from *C. capsularis* (white jute)(Zhang et al. 2015b). Biswas et al.(Biswas et al. 2015) constructed nine linkage groups, including 458 SNPs, using transcriptome sequence data. Tao et al. constructed a white jute genetic map including

11 linkage groups (LGs) and 913 polymorphic markers covering 1621.4 cM with an average density of 1.61 cM per locus (Tao et al. 2017).

Insertion/deletion (InDel) markers can easily be detected among accessions using whole-genome sequencing and high-throughput genetic analysis, and they may be one of the most abundant sources of DNA/RNA variation in plant genomes (Lv et al. 2013; Liu et al. 2016). At present, InDels, SSRs, and SNPs are mainstream and the most commonly used markers due to their efficiency, low cost, and simplicity in detection compared with other molecular markers. Additionally, with the development of NGS, InDels have been widely applied to genetic diversity analyses, genetic map construction, association mapping analysis, and marker-assisted selection (MAS) breeding in many crops (Rao et al. 2011; Lv et al. 2013). However, currently, development and application of InDels have not been reported for jute. To accelerate the application of more molecular markers in jute breeding and gene mapping we identified and developed jute InDel markers, for the first time, using three transcriptome datasets. The objectives were the following: (1) to estimate the frequencies and characteristics of InDels in jute transcriptomes, (2) to develop large-scale InDel markers through transcriptome comparative analysis, and (3) to verify PCR-based InDels and evaluate their application efficiency among 62 jute accessions. These results will provide new tools for jute molecular research and breeding.

Materials and methods

Plant material and genotyping

Twenty-six *C. capsularis* and thirty-six *C. olitorius* accessions (Table 1) were genotyped to evaluate InDel polymorphisms. Three of these accessions, two *C. olitorius* ('NY' and 'TC') and one *C. capsularis* ('Huangma 179'), have been also used for transcriptome sequencing. The *C. capsularis* accessions included 5 wild species and 21 non-wild accessions, and the *C. olitorius* accessions were composed of 4 wild species and 32 non-wild accessions.

Transcriptome de novo assembly and InDel polymorphisms screening

In total, 188.66 Gb clean read bases were assembled de novo to generate reference transcriptome sequences and screen InDel polymorphisms. The data were composed of two parts. One part was from sequences from a transcriptome sequence of the root and leaf tissues of two *C. olitorius* genotypes (NY and TC) in our previous study (NCBI accession ID no. SRP116874), and the other part was from RNA-seq data for 'Huangma 179' that were downloaded from the NCBI database (NCBI accession ID no. SRP060467). Transcriptome assembly was accomplished using the software program Trinity(Grabherr et al. 2011) with the parameter 'min_kmer_cov' set to 2 and all other parameters set to the default. Unigenes were annotated in all seven databases, National Center for Biotechnology Information (NCBI) non-redundant protein sequences (Nr); NCBI non-redundant nucleotide sequences (Nt); Protein Family (PFAM); Clusters of Orthologous Groups of Proteins (COG); Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG) Ortholog (KO) database; and Gene Ontology (GO) using the Basic Local Alignment Search Tool with

an E-value threshold of 10^{-5}) ((Zhang et al. 2015c). Picard-tools v1.41 and samtools v0.1.18(Li et al. 2009; Yang et al. 2017) were used to sort and remove duplicated reads and merge the bam alignment results of each sample based on reference transcriptome sequences. The GATK3 software(Van der Auwera et al. 2013) was used to perform InDel calling. Raw vcf files were filtered with GATK's standard filter method and other parameters (i.e., FS > 60.0, MQ < 40.0).

Development of InDel markers and polymorphism validation

To evaluate the rate of InDel loci predicted by computer successfully falling to experimental InDel marker, we designed 50 primer pairs based on InDel sites for PCR amplification of the 62 jute accessions using the Primer Premier 5.0 program(Ren et al. 2004). First, the primers were screened in the three jute accessions ('NY' and 'TC', and 'Huangma 179') to identify polymorphic primers and then the primers with polymorphisms were used in the genotyping and cluster analysis of the accessions. The predicted annealing temperature was about 55 °C, and primer length ranged from 18 to 26 bp with amplicon sizes of 100–500 bp. All primers were synthesized by Sangon Biotech (Shanghai) Co., Ltd. (Shanghai, China).

Genomic DNA for each accession was extracted using a DNeasy Plant Mini Kit (Tiangen, Tiangen Biotech Co.LTD. Beijing, China) and then diluted to the working concentration for InDel genotyping (approximately 20 ng/mL). DNA amplification was performed in 10 µL reaction volumes with 1 µL 10X PCR buffer, 0.3 mM dNTP, 0.15 µL (2U/µL) Taq DNA polymerase (Tiangen), 0.5 µL µmol/L each primer, 1.5 µL DNA, and 5.99 µL ddH₂O. The following PCR program was used: 5 min at 94 °C;

followed by 33 cycles of 30 s at 95 °C, 40 s at 55 °C, 60 s at 72 °C; and a final extension of 10 min at 72 °C. The PCR products were separated on 8% polyacrylamide gels with electrophoresis, and silver staining was carried out according to Zhang et al.(Zhang et al. 2000). The clear bands from PCR products were used for genotyping.

Genetic diversity was evaluated using the program PowerMarker V3.25(Liu and Muse 2005), which measured the number of alleles (na), gene diversity, and the polymorphic information content (PIC) for each locus. Furthermore, PowerMarker V3.25 was applied to calculate genetic distances among all accessions based on the methods used by Nei (Liu and Muse 2005), and a neighbor-joining (NJ) tree was constructed based on similarity measures. MEGA V6.0 software(Tamura et al. 2013) was used to display the NJ tree.

Results

Transcriptome assembly and unigene functional annotation

Clean reads were used to assemble 10 3281 non-redundant unigenes and the unigenes were used as reference sequences for downstream analysis. Unigene lengths varied between 201 and 15 541 nucleotides with average lengths of 1140 bp and N50 1758 bp(Fig. 1), respectively.67 487 (65.34%) unigenes were successfully annotated in at least one database. The most unigenes (58553; 56.69%) were annotated in the nr database, followed by SwissProt (48 138; 46.60%). (Fig. 2) (Table S1, S2).

InDel polymorphism discovery

In the current study, 51 172 InDel sites (Table S3, S4) in 18 800 unigenes were discovered among three jute accessions. The number of InDel sites in each unigene varied from 1 to 31 (Fig. 3). Unigenes with one InDel site were the most common (6800) followed by those with two InDel sites (4517), while there were only 31 unigenes with more than 15 InDel sites, which accounted for only 0.16% of all unigenes.

In total, 94 InDel types (defined by the InDel fragment length) were present in the three accessions. InDel types varied from 1 to 159 bp (Fig. 4). Moreover, The number of InDels decreased sharply as InDel length increased. The most common type was single-nucleotide (23 028) followed by binucleotide (9824) and tri-nucleotide (9182). These three types accounted for more than 79 % of all InDels. Furthermore, we compared the nucleotide composition of these three InDel types (Table S5) and found that for single-nucleotide types, InDels of A (8423) and T (8415) were overwhelmingly most common. For binucleotide InDels, there were four predominant pairs (TT, AA, TA and AT; 52.60 % of binucleotide InDels), and for trinucleotide InDels were four principal triplets (TAA, ATT, AAA and TTT; 16.50% of trinucleotide InDels). The A and T content of bi- and tri-nucleotide InDels was 71.52% and 65.82%, respectively.

In total, 49 563 polymorphism InDel sites in 18445 unigenes were discovered in the comparison between TC and YG, followed by 48 934 InDel sites in 18 408 unigenes between NY and YG, and 3570 InDels in 2701 unigenes between NY and

TC Additionally, of all possible pairs of accessions, there were 1273 InDel loci in 1129 unigenes with polymorphisms.

Analysis of polymorphisms in the InDel markers

In total, 40 of the 50 primer pairs successfully amplified bands, and 29 of these represented polymorphisms among the three jute accessions ('NY', 'TC', and 'Huangma 179') (Table S6). Genetic diversity of the polymorphic InDel loci was assessed for 62 accessions. The average number of alleles was 2.966. Seven InDels had two alleles, sixteen had three, and six had four. PIC varied from 0.340 to 0.680, with an average of 0.491. Average gene diversity was 0.575 and varied from 0.434 to 0.731. InDel-20 had the highest PIC (0.680) and genetic diversity (0.731) followed by InDel-18 (PIC = 0.665, genetic diversity = 0.715). InDel-28 had the lowest PIC (0.340) and genetic diversity (0.434), and the highest major allele frequency (0.682) (Table 2).

Evaluation of the relationship and clustering of 62 jute accessions

Based on the above results, 29 InDel primer pairs were used to evaluate the genetic relationship and differentiation of 62 *Corchorus* spp. accessions (26 *C. capsularis* and 36 *C. olitorius*). The accessions were unambiguously classified into two clusters (Fig. 5). A cluster included all of 26 *C. capsularis* accessions. And the other cluster included all of 36 *C. olitorius* accessions. And obviously, The wild type accessions (*C. olitorius* : Yuye-II, Maliyeshengchangguo, Nanyangyesheng and

DS/059C; *C. capsularis*: Zhisuma, Y/086C, Aidianyehuangma, YunyeI-4, Lianjianghuangma) were more closely related to each other than to the other types (cultivars, landraces, genetic materials).

Discussion

Although the jute genome (Islam et al. 2017) and a few transcriptomes (Zhang et al. 2015c) have been published in the last two years, molecular markers and high-density genetic maps are still lacking on jute compared with other crops. At present, the densest genetic map includes only 913 polymorphic markers (Tao et al. 2017), which greatly limits the development of gene clones and molecular marker-assisted selection. Currently, most molecular markers used in jute are RAPDs, SRAPs, ISSRs, and SSRs (Das et al. 2012; Zhang et al. 2015a; Zhang et al. 2016; Saha et al. 2017), and the number of markers based on PCR are limited. Further, the majority of jute molecular markers comprise 1,906 EST-SSR markers developed by Zhang et al., which is far from meeting the need for high-density genetic map construction and gene mapping for molecular markers. In addition, the use of the three molecular markers (ISSR, SRAP, and RAPD) has been limited in application because of the technological complexity of ISSR and SRAP and poor repeatability of RAPD. InDel markers can easily be developed by blasting genomes or transcriptomes of different accessions using next generation sequencing and bioinformatics (Liu et al. 2016). In this paper, we identified and developed InDel markers using the three transcriptome datasets for the first time, and the most of unigenes were annotated in one to seven

databases, aiding the development of functional molecular markers for gene mapping of important jute phenotypes.

Liu et al.(Wei et al. 2014; Liu et al. 2016) identified 25 701 potential InDel markers with 82.5 % small InDel fragments (1–20 bp) in the watermelon genome. Compared with the study mentioned above, more InDel sites were discovered in jute transcriptomes in the current study, indicating that InDel sites are abundant in jute. A higher rate (99.84%) of small InDel fragments (1–15 bp) was also discovered in this study, consistent with the other crops(Liu et al. 2013; Wei et al. 2014). In the current study, A and T were the overwhelming predominant nucleotides for single-nucleotide InDels, consistent with previous studies(Wei et al. 2014). More InDel sites were discovered among TC, NY, and YG than between TC and NY, which might be because TC and NY were both *C. olitorius*, while YG was *C. capsularis*. Of the 50 primers, 29 (58%) exhibited polymorphisms, similar to the amount of polymorphism exhibited in *Citrullus lanatus*(Liu et al. 2016). The PIC value varied from 0.340 to 0.680 with an average of 0.491 implying that most InDels were within a moderate range of genetic diversity, which is consistent with the other types of markers in jute(Grabherr et al. 2011; Rana et al. 2013; Ghosh et al. 2014). In addition, polymorphic InDel markers were successfully used in a cluster analysis of 62 jute accessions, which indicated that our main aim (the large-scale development of InDel markers) was successfully achieved. These markers will provide a valuable resource

for constructing a genetic linkage map and conducting QTL analyses and molecular marker-assisted selection breeding in jute.

Acknowledgements

This study was financially supported by the Hunan Natural Science Foundation (grant number 2017JJ3350), the National Natural Science Foundation of China (grant number 31601351), the Science and Technology Innovation Project of Chinese Academy of Agricultural Sciences (grant number ASTIP-IBFC01), the National Key Technology Support Program (grant number 2013BAD01B03-13) and the Germplasm Resources Protection Project (grant number 2016NWB044).

Reference

- Banerjee S, Das M, Mir RR, Kundu A, Topdar N, Sarkar D, Sinha MK, Balyan HS, Gupta PK. 2012. Assessment of Genetic Diversity and Population Structure in a Selected Germplasm Collection of 292 Jute Genotypes by Microsatellite (SSR) Markers. *Molecular Plant Breeding*.
- Biswas C, Dey P, Karmakar P, Satpathy S. 2015. Discovery of large-scale SNP markers and construction of linkage map in a RIL population of jute (*Corchorus capsularis*). *Molecular Breeding* 35(5): 119.
- Das M, Banerjee S, Dhariwal R, Vyas S, Mir RR, Topdar N, Kundu A, Khurana JP, Tyagi AK, Sarkar D. 2012. Development of SSR markers and construction of a linkage map in jute. *Journal of Genetics* 91(1): 21-31.
- Ghosh A, Sharmin S, Islam S, Pahloan MU, Islam S, Khan H. 2010. SSR markers linked to mite (*Polyphagotarsonemus latus* Banks) resistance in jute (*Corchorus olitorius* L.). *Czech J Genet Plant Breed* 46(2): 64-74.
- Ghosh RK, Wongkaew A, Sreewongchai T, Nakasathien S, Phumichai C. 2014. Assessment of Genetic Diversity and Population Structure in Jute (*Corchorus* spp.) Using Simple Sequence Repeat (SSR) and Amplified Fragment Length Polymorphism (AFLP) Markers. *Kasetsart Journal - Natural Science* 48(1): 83-94.
- Ghosh S, Meena K, Sinha MK, Karmakar PG. 2015. Genetic Diversity in *Corchorus olitorius* Genotypes Using Jute SSRs. *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences* 87(3): 917-926.

- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29(7): 644-652.
- Islam MS, Saito JA, Emdad EM, Ahmed B, Islam MM, Halim A, Hossen Q, Hossain MZ, Ahmed R, Hossain MS. 2017. Comparative genomics of two jute species and insight into fibre biogenesis. *Nature plants* 3: 16223-16223.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16): 2078-2079.
- Liu B, Wang Y, Zhai W, Deng J, Wang H, Cui Y, Cheng F, Wang X, Wu J. 2013. Development of InDel markers for *Brassica rapa* based on whole-genome re-sequencing. *Theoretical & Applied Genetics* 126(1): 231-239.
- Liu G, Xu J, Zhang M, Li P, Yao X, Hou Q, Zhu L, Ren R, Yang X. 2016. Exploiting Illumina sequencing for the development of InDel markers in watermelon (*Citrullus lanatus*). *The Journal of Horticultural Science and Biotechnology* 91(3): 220-226.
- Liu K, Muse SV. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21(9): 2128-2129.
- Lv H, Yang L, Kang J, Wang Q, Wang X, Fang Z, Liu Y, Zhuang M, Zhang Y, Lin Y. 2013. Development of InDel markers linked to *Fusarium* wilt resistance in *cabbage*. *Molecular breeding* 32(4): 961-967.

- Mir J, Roy A, Ghosh S, Karmakar P. 2011. Development of linkage map in F2 population of selected parents with respect to *Macrophomina phaseolina* resistance trait using screened polymorphic RAPD and developed SCAR markers of jute. *Archives of phytopathology and plant protection* 44(7): 671-683.
- Mir RR, Banerjee S, Das M, Gupta V, Tyagi AK, Sinha MK, Balyan HS, Gupta PK. 2009. Development and Characterization of Large-Scale Simple Sequence Repeats in Jute. *Crop Science* 49(5): 1687.
- Rana MK, Arora K, Singh S, Singh AK. 2013. Multi-locus DNA fingerprinting and genetic diversity in jute (*Corchorus* spp.) based on sequence-related amplified polymorphism. *Journal of Plant Biochemistry & Biotechnology* 22(1): 1-8.
- Rao IS, Srikanth B, Kishore VH, Suresh PB, Chaitanya U, Vemireddy L, Voleti S, Subbarao L, Rani NS, Sundaram R. 2011. Indel polymorphism in sugar translocation and transport genes associated with grain filling of rice (*Oryza sativa* L.). *Molecular breeding* 28(4): 683-691.
- Ren L, Zhu B-q, Zhang Y-b, Wang H-y, Li C-y, Su Y, Ba C-f. 2004. The research of applying primer premier 5.0 to design PCR primer. *J Jinzhou Med Coll* 25: 43-46.
- Saha D, Rana RS, Chakraborty S, Datta S, Kumar AA, Chakraborty AK, Karmakar PG. 2017. Development of a set of SSR markers for genetic polymorphism detection and interspecific hybrid jute breeding. *The Crop Journal*.
- Satya P, Paswan PK, Ghosh S, Majumdar S, Ali N. 2016. Confamilial transferability

of simple sequence repeat (SSR) markers from cotton (*Gossypium hirsutum* L.) and jute (*Corchorus olitorius* L.) to twenty two Malvaceous species. *3 Biotech* 6(1): 65.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution* 30(12): 2725-2729.

Tao A, Huang L, Wu G, Afshar RK, Qi J, Xu J, Fang P, Lin L, Zhang L, Lin P. 2017. High-density genetic map construction and QTLs identification for plant height in white jute (*Corchorus capsularis* L.) using specific locus amplified fragment (SLAF) sequencing. *BMC genomics* 18(1): 355.

Topdar N, Kundu A, Sinha MK, Sarkar D, Das M, Banerjee S, Kar CS, Satya P, Balyan HS, Mahapatra BS et al. 2013. A complete genetic linkage map and QTL analyses for bast fibre quality traits, yield and yield components in jute (*Corchorus olitorius* L.). *Cytology and Genetics* 47(3): 129-137.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J. 2013. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*: 11.10. 11-11.10. 33.

Wei L, Miao H, Li C, Duan Y, Niu J, Zhang T, Zhao Q, Zhang H. 2014. Development of SNP and InDel markers via de novo transcriptome assembly in *Sesamum indicum* L. *Molecular Breeding* 34(4): 2205-2217.

- Yang Q, Shohag M, Feng Y, He Z, Yang X. 2017. Transcriptome Comparison Reveals the Adaptive Evolution of Two Contrasting Ecotypes of Zn/Cd Hyperaccumulator *Sedum alfredii* Hance. *Frontiers in plant science* 8.
- Zhang J, Wu Y, Guo W, Zhang T. 2000. Fast screening of microsatellite markers in cotton with PAGE/silver staining. *Acta Gossypii Sinica* 12(5): 267-269.
- Zhang L, Cai R, Yuan M, Tao A, Xu J, Lin L, Fang P, Qi J. 2015a. Genetic diversity and DNA fingerprinting in jute (*Corchorus* spp.) based on SSR markers. *The Crop Journal* 3(5): 416-422.
- Zhang L, Li Y, Tao A, Fang P, Qi J. 2015b. Development and characterization of 1,906 EST-SSR markers from unigenes in jute (*Corchorus* spp.). *PloS one* 10(10): e0140861.
- Zhang L, Liu X, Zhang L, Wan X, Tao A, Fang P, Lin P, Qi J. 2016. Construction of a genetic map using newly developed SSR markers for identifying QTL for plant height in jute (*Corchorus capsularis*). Elsevier.
- Zhang L, Ming R, Zhang J, Tao A, Fang P, Qi J. 2015c. De novo transcriptome sequence and identification of major bast-related genes involved in cellulose biosynthesis in jute (*Corchorus capsularis* L.). *BMC genomics* 16(1): 1062.

Tables

Table 1. Information for 26 *Corchorus capsularis* L. and 36 *Corchorus olitorius* L. accessions used for InDel genotyping.

Accession no.	Name	Type	Species	Accession no.	Name	Type	Species
1	IJ019 号	cultivars	<i>C. capsularis</i> L	32	y003-2	unknown	<i>C. olitorius</i> L.
2	Xinye	cultivars	<i>C. capsularis</i> L	33	IV-Hig	landraces	<i>C. olitorius</i> L.
3	JG010	cultivars	<i>C. capsularis</i> L	34	Pama72-1	cultivars	<i>C. olitorius</i> L.
4	Changwuhma	cultivars	<i>C. capsularis</i> L	35	2002-36	genetic materials	<i>C. olitorius</i> L.
5	Jialiqingjing	cultivars	<i>C. capsularis</i> L	36	1-23	landraces	<i>C. olitorius</i> L.
6	Taima9	landraces	<i>C. capsularis</i> L	37	78-1077	genetic materials	<i>C. olitorius</i> L.
7	c2002-11	genetic materials	<i>C. capsularis</i> L	38	2002-049	genetic materials	<i>C. olitorius</i> L.
8	Hongzhongma	landraces	<i>C. capsularis</i> L	39	K42	landraces	<i>C. olitorius</i> L.
9	Hezhi20hao	cultivars	<i>C. capsularis</i> L	40	Xiatang	landraces	<i>C. olitorius</i> L.
10	I40402-8	cultivars	<i>C. capsularis</i> L	41	Co24	unknown	<i>C. olitorius</i> L.
11	Meifengma	cultivars	<i>C. capsularis</i> L	42	Y008	unknown	<i>C. olitorius</i> L.
12	72-01	cultivars	<i>C. capsularis</i> L	43	SM/028CG	unknown	<i>C. olitorius</i> L.
13	Weiyuanma	cultivars	<i>C. capsularis</i> L	44	Zhejiangma	landraces	<i>C. olitorius</i> L.
14	Minma369	cultivars	<i>C. capsularis</i> L	45	Y022-7	unknown	<i>C. olitorius</i> L.
15	lingsheng	cultivars	<i>C. capsularis</i> L	46	Tianma	landraces	<i>C. olitorius</i> L.
16	Zhongyuanma	landraces	<i>C. capsularis</i> L	47	NY	unknown	<i>C. olitorius</i> L.
17	ACC-N0-4681	genetic materials	<i>C. capsularis</i> L	48	TC	unknown	<i>C. olitorius</i> L.
18	Tengying4	landraces	<i>C. capsularis</i> L	49	Maliyeshengchangguo	wild germplasm	<i>C. olitorius</i> L.
19	Zhejiianglinpin	landraces	<i>C. capsularis</i> L	50	Yunye-II	wild germplasm	<i>C. olitorius</i> L.
20	Gengxingma	unknown	<i>C. capsularis</i> L	51	Nanyangyesheng	wild germplasm	<i>C. olitorius</i> L.

21	Zhisuma	wild germplasm	<i>C. capsularis</i> L	52	Kuanyechangguo	cultivars	<i>C. olitorius</i> L.
22	Y/086C	wild germplasm	<i>C. capsularis</i> L	53	DS/059C	wild germplasm	<i>C. olitorius</i> L.
23	Aidianyehuangma	wild germplasm	<i>C. capsularis</i> L	54	changjia17	unknown	<i>C. olitorius</i> L.
24	YunyeI-4	wild germplasm	<i>C. capsularis</i> L	55	DS/038C	unknown	<i>C. olitorius</i> L.
25	Lianjianghuangma	wild germplasm	<i>C. capsularis</i> L	56	JRC/558	unknown	<i>C. olitorius</i> L.
26	Yuanguo179	cultivars	<i>C. capsularis</i> L	57	BL/106CC	unknown	<i>C. olitorius</i> L.
27	III84	genetic materials	<i>C. olitorius</i> L.	58	SM/056	unknown	<i>C. olitorius</i> L.
28	090-8	genetic materials	<i>C. olitorius</i> L.	59	K-58	genetic materials	<i>C. olitorius</i> L.
29	090-11	genetic materials	<i>C. olitorius</i> L.	60	BL/013	genetic materials	<i>C. olitorius</i> L.
30	y05-05	unknown	<i>C. olitorius</i> L.	61	DS/068CR	genetic materials	<i>C. olitorius</i> L.
31	JRC/580C	unknown	<i>C. olitorius</i> L.	62	ACC. NO. 4792	genetic materials	<i>C. olitorius</i> L.

Note: cultivars: widely cultivated varieties developed through various breeding methods; landraces which were selected and developed by local farmers and planted chronically only in certain places; genetics materials, referred to number or structure variation of chromosomes or genes with particular value which could be transmitted to offspring and included intermediate breeding lines, such as sterile line, aneuploidy, substitution line, etc.

Table 2. Genetic diversity information for 29 polymorphic InDel loci in 62 jute (*Corchorus* spp.) accessions.

Marker	Major Allele Frquency	The number of Allele	Gene diversity	PIC
InDel-1	0.571	4.000	0.576	0.510
InDel-2	0.631	3.000	0.519	0.451
InDel-3	0.597	3.000	0.533	0.453
InDel-4	0.468	3.000	0.638	0.566
InDel-5	0.375	3.000	0.664	0.590
InDel-6	0.433	3.000	0.649	0.576
InDel-7	0.525	3.000	0.514	0.398
InDel-8	0.517	3.000	0.556	0.460
InDel-9	0.500	3.000	0.568	0.474
InDel-10	0.516	2.000	0.499	0.375
InDel-11	0.613	3.000	0.545	0.482
InDel-12	0.509	3.000	0.590	0.508
InDel-13	0.581	2.000	0.487	0.368
InDel-14	0.385	4.000	0.696	0.637
InDel-15	0.452	4.000	0.688	0.638
InDel-16	0.597	3.000	0.552	0.484
InDel-17	0.500	3.000	0.622	0.550
InDel-18	0.404	4.000	0.715	0.665
InDel-19	0.532	4.000	0.599	0.529
InDel-20	0.302	4.000	0.731	0.680
InDel-21	0.452	3.000	0.594	0.506
InDel-22	0.500	2.000	0.500	0.375
InDel-23	0.545	3.000	0.579	0.502
InDel-24	0.588	2.000	0.484	0.367
InDel-25	0.581	3.000	0.533	0.444
InDel-26	0.484	3.000	0.631	0.559
InDel-27	0.581	2.000	0.487	0.368
InDel-28	0.682	2.000	0.434	0.340
InDel-29	0.545	2.000	0.496	0.373
Mean	0.516	2.966	0.575	0.491

Figure captions

Fig. 1. Frequency analysis of lengths of all unigenes.

Fig. 2. Unigene annotation success rates across multiple databases.

Fig. 3. Frequency of InDels per unigene for three jute (*Corchorus* spp.) accessions.

Fig. 4. Frequency of InDel types (based on InDel length) for three accessions of jute (*Corchorus* spp.)

Fig. 5. The 62 accessions were unambiguously classified into two clusters. A cluster included all of 26 *C. capsularis* accessions. And the other cluster included all of 36 *C. olitorius* accessions. Non wild accessions included cultivars, landraces, genetic materials and unknown type.

Draft

Fig. 1. Frequency analysis of lengths of all unigenes

1524x1143mm (100 x 100 DPI)

Fig. 2. Unigene annotation success rates across multiple databases.

102x94mm (300 x 300 DPI)

Fig. 3. Frequency of InDels per unigene for three jute (*Corchorus* spp.) accessions.

121x91mm (300 x 300 DPI)

Fig. 4. Frequency of InDel types (based on InDel length) for three accessions of jute (*Corchorus* spp.)

289x202mm (150 x 150 DPI)

Fig. 5. The 62 accessions were unambiguously classified into two clusters. A cluster included all of 26 *C. capsularis* accessions. And the other cluster included all of 36 *C. olitorius* accessions. Non wild accessions included cultivars, landraces, genetic materials and unknown type.

1763x1322mm (72 x 72 DPI)