Tan et al. BMC Plant Biology (2019) 19:517
https://doi.org/10.1186/s12870-019-2133-z

BMC Plant Biology

RESEARCH ARTICLE Open Access

Single-molecule real-time sequencing facilitates the analysis of transcripts and splice isoforms of anthers in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*)

Chong Tan†, Hongxin Liu†, Jie Ren, Xueling Ye, Hui Feng and Zhiyong Liu*

Abstract

Background: Anther development has been extensively studied at the transcriptional level, but a systematic analysis of full-length transcripts on a genome-wide scale has not yet been published. Here, the Pacific Biosciences (PacBio) Sequel platform and next-generation sequencing (NGS) technology were combined to generate full-length sequences and completed structures of transcripts in anthers of Chinese cabbage.

Results: Using single-molecule real-time sequencing (SMRT), a total of 1,098,119 circular consensus sequences (CCSs) were generated with a mean length of 2664 bp. More than 75% of the CCSs were considered full-length non-chimeric (FLNC) reads. After error correction, 725,731 high-quality FLNC reads were estimated to carry 51,501 isoforms from 19,503 loci, consisting of 38,992 novel isoforms from known genes and 3691 novel isoforms from novel genes. Of the novel isoforms, we identified 407 long non-coding RNAs (lncRNAs) and 37,549 open reading frames (ORFs). Furthermore, a total of 453,270 alternative splicing (AS) events were identified and the majority of AS models in anther were determined to be approximate exon skipping (XSKIP) events. Of the key genes regulated during anther development, AS events were mainly identified in the genes *SERK1*, *CALS5*, *NEF1*, and *CESA1/3*. Additionally, we identified 104 fusion transcripts and 5806 genes that had alternative polyadenylation (APA).

Conclusions: Our work demonstrated the transcriptome diversity and complexity of anther development in Chinese cabbage. The findings provide a basis for further genome annotation and transcriptome research in Chinese cabbage.

Keywords: Chinese cabbage, Anther, Full-length transcript, Alternative splicing, Fusion transcript

Background

Gene sequencing emerged as a revolutionary technology in the field of biological research. The first of these technologies was Sanger sequencing; however, due to low throughput and poor automation, Sanger sequencing was severely limited in its application in genome and transcriptome analysis [1]. The advent of NGS technologies, such as ABI SOLiD, Illumina Solexa, and Roche 454 systems, stimulated structural and functional genomics studies for diverse plant species. Among these technologies, Illumina sequencing has the advantages of high accuracy, high throughput, high sensitivity, and low cost, and is now the most widely used platform in genome sequencing [2]. *C. sativus* was the first vegetable crop to complete genome-wide de novo sequencing by NGS. Subsequently, the main crop genomes of *S. tuberosum*, *T. aestivum*, *B. napus*, *G. raimondii*, and other crops were sequenced. Short-read RNA-Seq by NGS is frequently applied for transcriptome analysis. Using short-read RNA-Seq, researchers can obtain profiles for genome-wide expressed genes, including low-abundance genes, as well as new genes and SNPs [3]. Research on gene expression profiling of pollen and anther development in the genus *Brassica* has accumulated in recent years [4–11]. However,

* Correspondence: 2010500026@syau.edu.cn
†Chong Tan and Hongxin Liu contributed equally to this work.
College of Horticulture, Shenyang Agricultural University, Shenyang, Liaoning 110866, People's Republic of China

although NGS technologies are effective, they still have several drawbacks, including the generation of relatively short reads, which may lead to misassembly and gaps [12]. Moreover, short reads are not well suited to accurately detecting structural variations (SVs) and transcript isoforms generated by AS events [13, 14]. Limited by NGS methods, short RNA-Seq reads must be assembled into longer DNA contigs [15], a process that is susceptible to misassembly of short sequence reads transcribed from highly repetitive regions or similar members of multiple gene families [16]. This problem may become even more severe for polyploid plants that often harbor higher sequence similarity between coexisting subgenomes, which frequently indirectly leads to annotation error. Moreover, short-read RNA-Seq cannot distinguish between alternatively spliced forms for individual transcripts, which can make up a large proportion of transcripts. For instance, approximately 83.4% of multiple-exon genes are subject to AS in *A. thaliana*, which contributes to organismal protein diversity without massively increasing the number of genes [17].

Third generation sequencing (TGS) technologies have recently been developed, which is known for single-molecule sequencing (SGS) and sequencing in real-time [18]. The first TGS technology platform was delivered by Helicos Biosciences, but it proved unworkable from the market because it was relatively slow, expensive, and generated short reads (~ 32 bp) [19]. Soon after, single-molecule real-time sequencing (SMRT) sequencing by PacBio emerged as unique opportunity for constructing full-length transcripts [20]. The distinguishing features of SMRT technology is the production of long reads. Initially, the average length of reads generated by SMRT technology was just ~ 1.5 kb, but is now 10-15 kb [21]. Therefore, SMRT can improve the accuracy of gene models as it allows generation of reads that cover full-length transcripts [14]. However, SMRT sequencing still has major technical defects and limitations, namely its relatively high cost, lower throughput, and high error rate. Therefore, at present, a combination of NGS technologies and SMRT sequencing is preferable: consensus sequence reads are constructed from raw PacBio subreads and aligned with the reads generated from appropriate NGS platforms. Using this approach, multiple complex genomes have been successfully de novo assembled or improved [22–30].

SMRT sequencing has been previously effectively applied to transcriptome analysis. Well-characterized full-length transcripts are not only beneficial for analysis of gene structure and alternative splicing, but also greatly improve functional studies of important loci [15]. Early applications of SMRT sequencing on the transcriptome were relatively narrow, and most focused on model organisms such as humans [13] and yeast [31]. Since 2015, SMRT technology has been widely applied to characterize the full-length sequence of genome and transcripts in diverse species. SMRT has facilitated structural genomics and grain transcriptome research in common hexaploid wheat [15]. In danshen, the application of SMRT sequencing to different root tissues revealed that about 40% of the detected gene loci had occurred alternative splicing (AS) events [32]. In maize B73, over 111,000 transcripts from six tissues were identified, unveiling the complexity of the transcriptome by SMRT sequencing [14]. PacBio SMRT was employed for the sorghum transcriptome, and over 11,000 novel splice isoforms, alternative polyadenylation (APA) of ~ 11,000 expressed genes, and more than 2100 novel genes were uncovered at an unprecedented scale [33]. The *A. thaliana* transcriptome was analyzed by SMRT, enhancing the understanding of differentially expressed AS isoforms under normal conditions and in response to ABA treatment [17]. In moso bamboo, over 42, 280 distinct splicing isoforms and 25,069 polyadenylation sites were found [34]. In *Dendrobium officinale*, the full-length cDNA transcripts of stems and leaves uncovered multiple genes involved in polysaccharide synthesis [35]. The red clover transcriptome was analyzed by SMRT sequencing and the results uncovered about 29,730 novel isoforms from known genes and 2194 novel isoforms from novel genes, in addition to over 5000 AS events, over 4300 long non-coding RNAs (lncRNAs), and 3700 fusion transcripts [36]. Using SMRT technology, a total of 113,321 transcripts were obtained from alfalfa leaves from three different development stages; sequencing data uncovered about 7568 AS events and 17,740 lncRNAs [37]. The above works are crucial for providing deeply understanding of their genome and transcripts.

The genus *Brassica* comprises multiple economically important vegetable and oil crops. The 'triangle of U' is well established and refers to three diploid species, *B. rapa* (A genome, 2n = 20), *B. nigra* (B genome, 2n = 16), and *B. oleracea* (C genome, 2n = 18), as well as three amphidiploid species, *B. napus* (AC genome, 2n = 38), *B. juncea* (AB genome, 2n = 36), and *B. carinata* (BC genome, 2n = 34). In 2011, the first genus *Brassica* genome draft, *B. rapa* genome v1.5, was published. The 283.8 Mb genome was generated using next-generation sequencing (NGS) technology with a contig N50 size of 46 kb, and greatly facilitated genomics and molecular biology research, as well as the generic breeding of *B. rapa* and other *Brassica* species [38]. The second version (v2.0) was assembled in 2017. Further improving the scaffold order, the upgraded *B. rapa* genome v2.5 was 389.2 Mb with a contig N50 size of 53 kb [39]. However, restricted by the read length of NGS technology, the above genome versions had the disadvantages of poor continuity, assembly errors, and low assembly rate of repetitive sequences. A more recent release, *B. rapa* genome v3.0 was de novo assembled and re-annotated using single-molecule sequencing (PacBio), optical mapping (BioNano), and chromosome conformation capture (Hi-

C) technologies. The total length of *B. rapa* genome v3.0 was 353.14 Mb, with a contig N50 size of 1.45 Mb and a scaffold N50 size of 4.45 Mb, including 1301 scaffolds and 389 gaps [40]. The high-quality reference genomic information lays a solid foundation for the development of genetics and functional genomics of *B. rapa*, especially the cloning of important agronomic trait regulatory genes and the analysis of genetic background. Only by analyzing the molecular mechanism of trait formation at the genetic level, can we carried out targeted genetic breeding, molecular marker-assisted breeding and even molecular design breeding, greatly improve breeding efficiency and accelerate the cultivation of excellent new varieties.

Owing to broad adaptability and numerous end-uses, Chinese cabbage is the most widely cultivated and consumed vegetable within *B. rapa*. Although the reference genome has been improved using the PacBio Sequel platform, sequence and structural data of tissue-specific mRNA remains scarce in Chinese cabbage. The main objective of this study was to characterize full-length transcripts in Chinese cabbage anthers using the emerging SMRT sequencing technology to unveil the transcriptome complexity of anther development. SMRT sequencing data, corrected by short-read NGS technology, were used to analyze full-length transcripts in anthers to further reveal AS events, lncRNAs, and fusion isoforms in Chinese cabbage. This study provides a valuable resource for further genome re-annotation and increases our understanding of the anther transcriptome.

Results

Transcriptome sequencing and error correction

Limited by the capacity of short-read RNA-Seq on an Illumina platform, anther-specific transcriptome analysis of Chinese cabbage double haploid (DH) line 'FT' (Fig. 1 a-c) was carried out using the PacBio Sequel platform. To identify the transcripts as completely as possible, high-quality total mRNAs from each of the pooled samples obtained throughout anther development were extracted and mixed to obtain full-length sequences and splice variants. The entire flow is shown in Fig. 2a.

Three different SMRT bell libraries were constructed and sequenced using the PacBio Sequel platform with cDNA insert sizes 1–2 kb, 2–3 kb, and > 3 kb. After filtering, a total of 1,895,346 polymerase reads representing more than 33.14 G bases were captured, with a mean length of 17,965 bp and N50 of 39,750 bp (Additional file 2: Table S1; Fig. 3a-c). After removing the adapter from polymerase reads, approximately 16,458, 266 filtered subreads were obtained with a mean length of 2121 bp (Additional file 2: Table S2). A total of 1,098, 119 circular consensus sequences (CCSs) with an average depth of 11.33 passes in three libraries were generated from subreads after merging and error correction by multiple sequencing (Additional file 2: Table S3). The length distribution of CCSs was consistent with the expected size of the three libraries (Fig. 3d-f). CCSs were counted separately as follows: 5′ primer, 3′ primer, poly-A tail, full-length, and full-length non-chimeric (FLNC). In total, we detected 863,281 full-length reads, containing the 5′ primer, 3′ primer, and poly-A tail. Then, 827,322 reads were considered to be FLNC with low artificial concatemers, accounting for 75.34% of CCSs (Table 1). The mean length of FLNC reads in the 1-2 K, 2-3 K, and > 3 K libraries were 1499 bp, 2324 bp, and 3288 bp, respectively (Fig. 3g-i; Table 1). The quantity distribution of FLNC reads in each library was similar, although the 1–2 kb library was slightly larger than the other two libraries. Overall, we have comprehensively obtained full-length transcripts, making it possible to accurately construct splice variants.

SMRT sequences have a high base error rate—up to 12–15%—mainly due to the extra insertion of bases. To further correct the FLNC reads sequenced by the PacBio Sequel platform, Illumina HiSeq 2000 transcripts of the anthers were employed by the proovread software. Using GMAP[2] software, the FLNC reads before and after error correction were compared to the reference genome for counting global and local percentage-of-identity (PID)

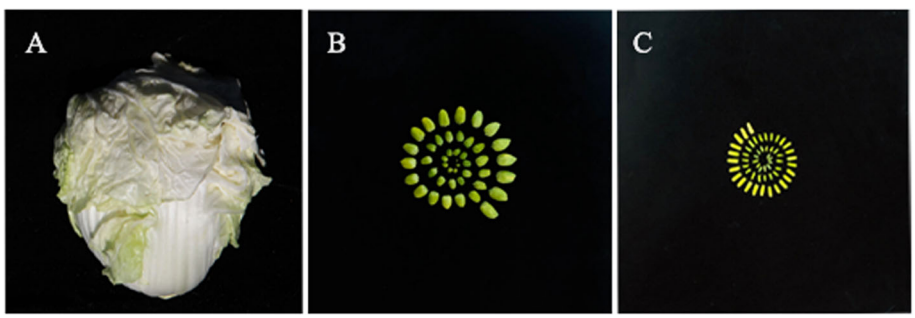

Fig. 1 Morphological characteristics of DH line 'FT'. **a** Leafy head. **b** The entire buds of inflorescence. **c** Anthers during different development stages

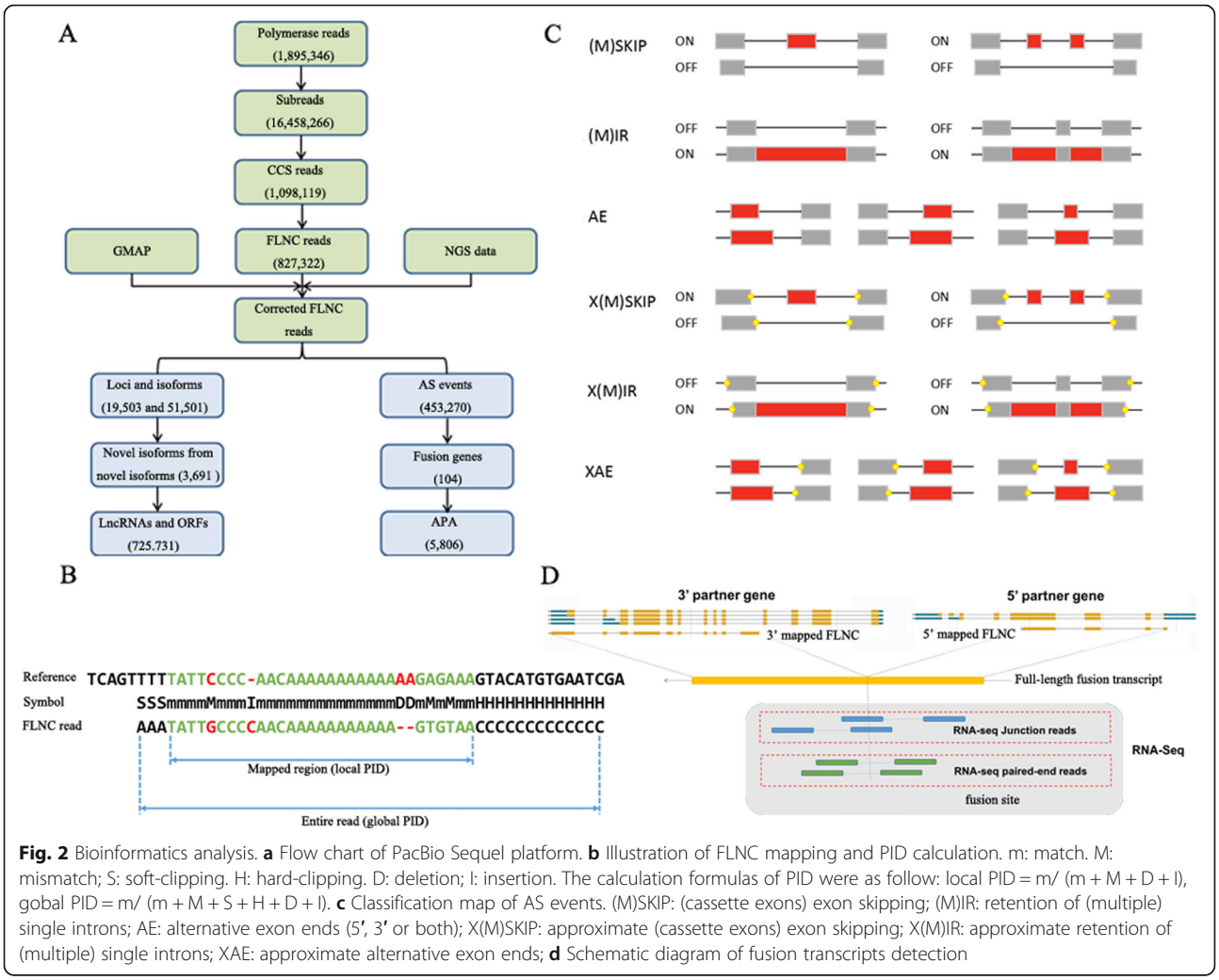

Fig. 2 Bioinformatics analysis. **a** Flow chart of PacBio Sequel platform. **b** Illustration of FLNC mapping and PID calculation. m: match. M: mismatch; S: soft-clipping. H: hard-clipping. D: deletion; I: insertion. The calculation formulas of PID were as follow: local PID = m/ (m + M + D + I), gobal PID = m/ (m + M + S + H + D + I). **c** Classification map of AS events. (M)SKIP: (cassette exons) exon skipping; (M)IR: retention of (multiple) single introns; AE: alternative exon ends (5′, 3′ or both); X(M)SKIP: approximate (cassette exons) exon skipping; X(M)IR: approximate retention of (multiple) single introns; XAE: approximate alternative exon ends; **d** Schematic diagram of fusion transcripts detection

(Fig. 4). Before error correction, the mean global PID was 94.97%. After error correction, the value was up to 97.04% (Additional file 2: Table S4). After updating, we obtained 725,731 high-quality FLNC reads for subsequent investigation (Table 2).

Loci and isoform detection and characterization

Error correction analysis allows accurate mapping of FLNC reads to the reference genome, including start site, termination site, and splicing site. Based on this information, gene loci and isoform can be identified. To assess isoform length density, we compared the loci coverage of the PacBio data set with the *B. rapa* v3.0 annotation. In our data, a total of 725,731 error-corrected FLNC reads covered 51,501 isoforms and were allocated to 19,503 loci. About 9102 loci were 1–2 kb in length, followed by 2–3 kb (3867), > 3 kb (3355), and < 1 kb (3179). In the reference genome, about 46,250 isoforms covered 46,250 loci, and the loci were mostly distributed at < 1 kb (24,937), followed by 1–2 kb (15,700), 2–3 kb (3959), and > 3 kb (1654) (Table 3; Fig. 5a). Similarly, we evaluated the isoform number of loci density, indicating that each locus could produce a unique isoform in the reference genome. However, in our data, approximately 12,124 (62.16%) loci could produce a unique isoform and more than five isoforms covered about 6.83% of the PacBio annotation loci (Fig. 5b). The gene A06.1469 had the largest number of isoforms at ~ 524. Thus, the PacBio data set could provide richer isoform length diversity and isoform number of loci density than the reference genome, which helped more fully reveal the complexity of the anther transcriptome. In addition, we evaluated the exon-intron structure of each loci and isoform obtained by PacBio Sequel platform. Among the 19,503 loci, there were 2911 (14.93%) single-exon loci and 16,592 (85.07%) multiple exon loci. Out of 51,501 isoforms, 4.188 (8.13%) were single exon, and 47, 313 (91.87%) were multiple exon (Table 3).

Based on the characteristics of library construction, we could not guarantee the structural integrity of the 5′ end of the transcripts. Therefore, the full-length evaluation

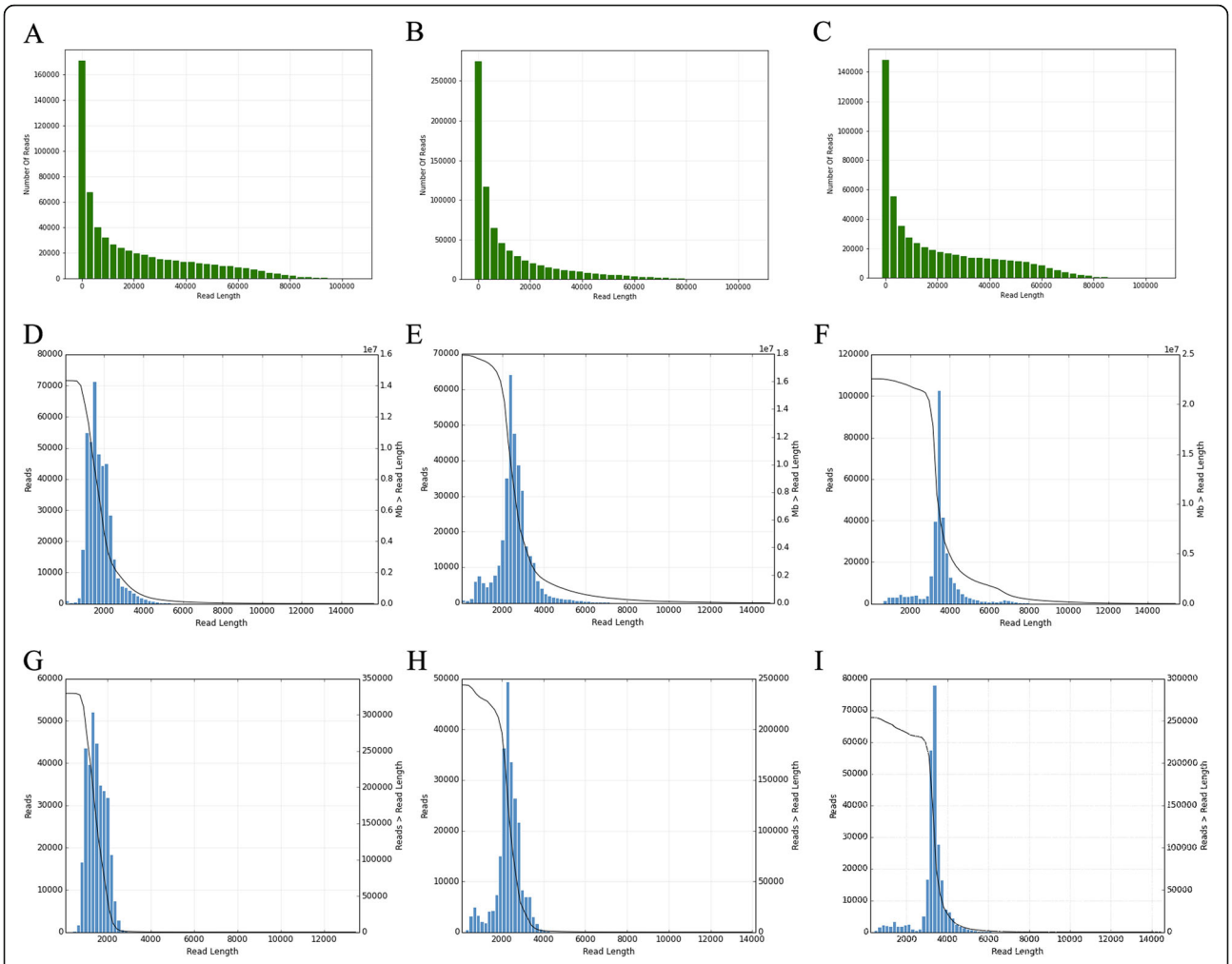

Fig. 3 Length distribution of the PacBio Sequel data output. **a-c** Number and length distribution of polymerase reads. **d-f** Number and length distribution of CCSs. **g-i** Number and length distribution of FLNC reads

of FLNC and isoforms produced by PacBio Sequel platform were only estimated at the 5′ end. With multiple-exon transcripts of genome annotation as a reference, isoforms obtained from the PacBio data set with identical direction and overlap greater than 20% were screened. If the first splice donor site at 5′ end of genome annotation was indeed included at the first splice donor site of the isoform obtained from PacBio data set, then the isoform was considered to be a full-length isoform, and the corresponding FLNC was considered to be a full-length FLNC. Our data indicated that approximately 76.66% multiple-exon isoforms and 88.22% multiple-exon FLNC contained the same splice donor site at the 5′ end as the reference annotation, and were regarded as full length, implying a relatively high integrity in structure (Additional file 2: Table S5).

Next, the sequenced gene loci and isoforms were compared with the reference annotation to determine novel loci or novel isoforms. The published *B. rapa* genome annotation contains 46,250 loci with 46,250 isoforms. In our PacBio data set, out of 51,501 isoforms from 19,503 genes, we identified 16,821 known isoforms from known

Table 1 Summary of ROI from the PacBio Sequel platform

Sample	Library	Cell	CCS	5′ primer	3′ primer	Poly-A	Full-length	FLNC	Mean FLNC length (bp)
Ant 1-2 k	1-2 K	D01	418,042	390,507	386,739	365,528.00	352,741	329,694	1499
Ant 2-3 K	2-3 K	E01	355,765	296,959	290,222	270,230.00	246,734	243,747	2324
Ant > 3 K	3 K+	A01	324,312	298,840	295,242	279,866.00	263,806	253,881	3288
Total	–	–	1,098,119	986,306	972,203	915,624.00	863,281	827,322	–

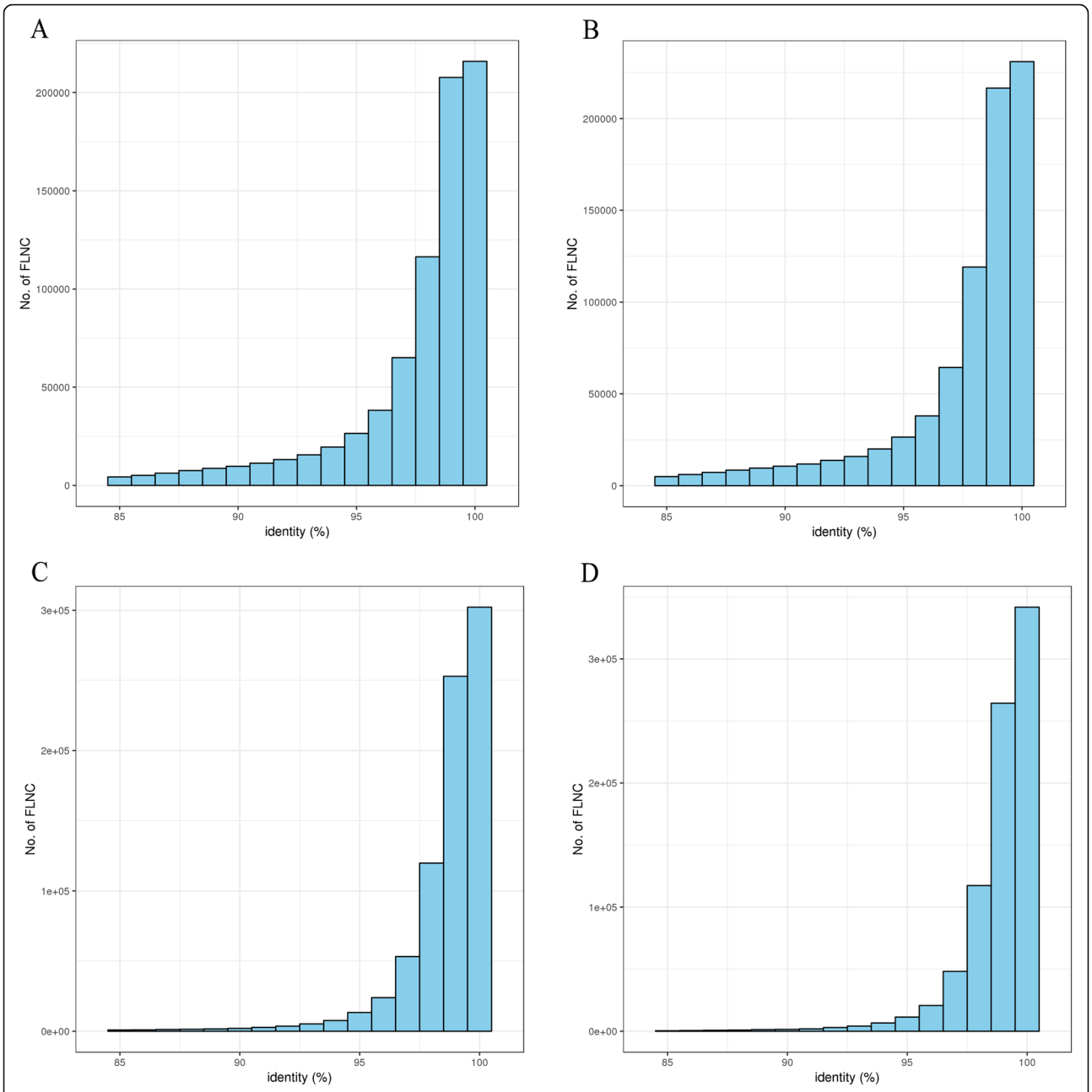

Fig. 4 The distribution of PID (percentage-of-identity) before and after error correction. **a** Gobal PID distribution before error correction. **b** Local PID distribution before error correction. **c** Gobal PID distribution after error correction. **d** Local PID distribution after error correction

Table 2 Classification of reference genome comparison results

Type	Pre-correction	Post-corrrction	Merge
Unmapped	3146 (0.38%)	616 (0.07%)	558 (0.07%)
Multiple-best map	7212 (0.87%)	7550 (0.91%)	6782 (0.82%)
Low PID map	230,229 (27.83%)	99,570 (12.04%)	94,251 (11.39%)
High quality map	586,735 (70.92%)	719,586 (86.98%)	725,731 (87.72%)

Table 3 Gene structure annotation

Type	Loci < 1 K	Loci 1-2 K	Loci 2-3 K loci	Loci > = 3 K	Total Loci	Single-exon loci	Multiple-exon loci	Total isoform	Single-exon isoform	Mutiple-exon isoform
Reference annotation	24,937 (53.92%)	15,700 (33.95%)	3959 (8.56%)	1654 (3.58%)	46,250	11,102 (24.00%)	35,148 (76.00%)	46,250	11,102 (24.00%)	35,148 (76.00%)
PacBio data set	3179 (16.30%)	9102 (46.67%)	3867 (19.83%)	3355 (17.20%)	19,503	2911 (14.93%)	16,592 (85.07%)	51,501	4188 (8.13%)	47,313 (91.87%)

genes. In addition, 2682 transcripts overlapping with no annotated gene were considered likely to be novel genes (Additional file 2: Table S6; Fig. 6d). Those novel genes were found to generate 3691 novel isoforms (Fig. 6c). We also found 38,992 novel isoforms from 11,398 known genes. Of the 3691 novel isoforms, 1455 (39.42%) were single-exon isoforms and 2236 (60.58%) were multiple-exon isoforms. The above novel genes and isoforms were beneficial to the improvement of the integrity of the *B. rapa* genome annotation.

Functional annotation of novel isoforms

In this study, all 3691 novel isoforms were functionally annotated by searching NCBI non-redundant protein sequences (NR) (89.76%), Gene Ontology (GO) (44.57%), EuKaryotic Orthologous Groups (KO) (24.36%), euKaryotic Ortholog Group (KOG) (22.70%), and Swiss-Prot Protein Sequence (Swiss-Prot) databases (42.81%), and a total of 377 (10.21%) were unannotated (Additional file 2: Table S7). A total of 420 novel isoforms had significant hits in all five databases (Fig. 7a). In the NR database, the largest three groups of novel isoforms were distributed in *B. rapa* (1578), *B. napus* (1236), and *B. oleracea* (112) (Fig. 7b). GO analysis assigned the enrichment of 1645 isoforms to three ontologies, namely, biological process, cellular component, and molecular function. We found 1602 GO terms in "biological process," of which "cellular process" (48.75%), "metabolic process" (45.53%), and "single-organism process" (32.52%) accounted for a large proportion. Many of the terms in "biological process" were associated with anther development, such as pollen germination, fatty acid metabolic process, pollen exine formation, pollen tube development, anther dehiscence, pollination, and pollen-postil interaction. A total of 358 GO terms were detected in "cellular component," with "cell" (45.40%), "cell part" (45.40%), and "membrane" (39.51%) the largest three enrichment terms. Eleven and two novel isoforms could be assigned to the GO terms "pollen tube tip" and "pollen tube," respectively in "cellular component." Our data showed that 769 GO terms were assigned to "molecular function", and the most highly abundant terms were "binding" (53.98%) and "catalytic activity" (47.29%) (Fig. 7c). To identify the enrichment pathways, a total 899 novel isoforms were subjected to 101 KEGG pathways. Novel isoforms in KEGG pathways consisted of five hierarchy: "cellular processes", "environmental information processing", "genetic information processing", "metabolism and "organismal systems". Among these terms, the most abundant hierarchy was "metabolism" (71.29%), followed by "genetic information processing" (11.88%) (Fig. 7d). In addition, we found three metabolic pathways associated with "fatty acid", which were essential for anther development [41]. KOG analysis shown that 838 novel isoforms were assigned to 24 categories, and the largest three classes

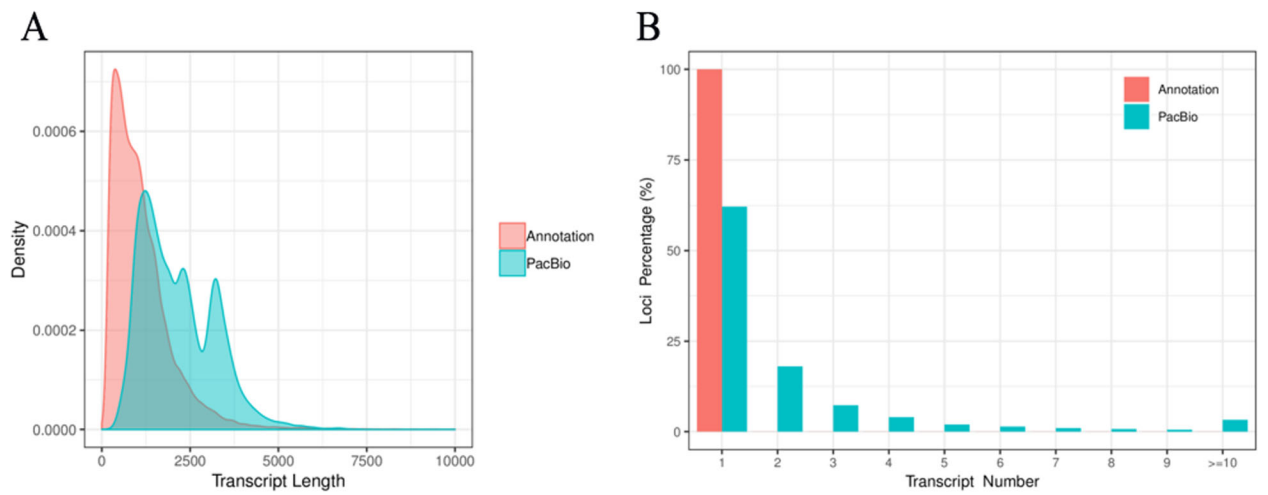

Fig. 5 Isoform length density and Isoform number of loci density. **a** The length distribution of all isoforms in the PacBio Sequel platform compared to the reference genome. **b** The number distribution of isoforms from each locus in the PacBio Sequel platform compared to the reference genome

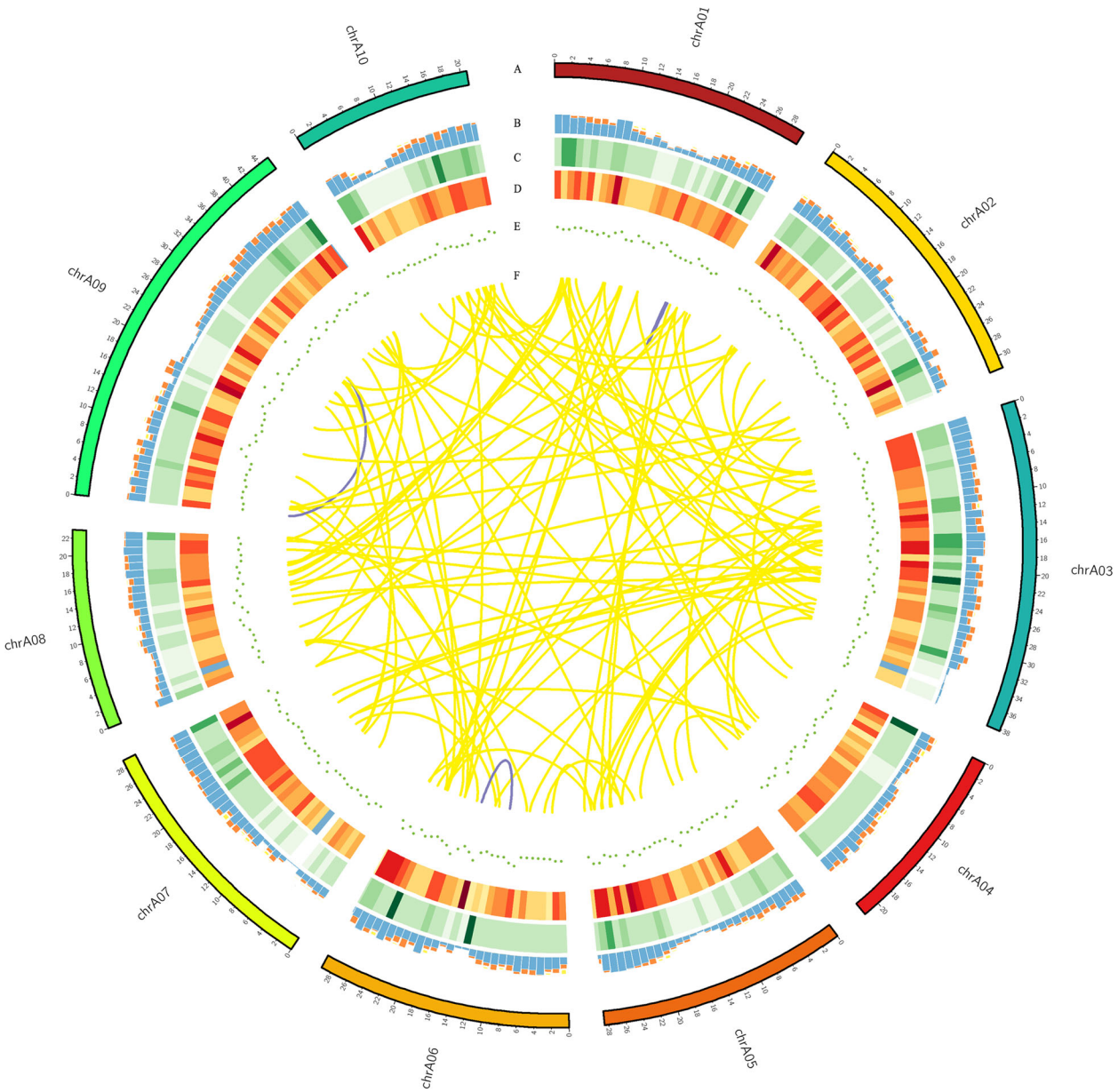

Fig. 6 Circos visualization of the PacBio Sequel platform at genome-wide level. **a** Ten chromosomes distribution of *B.rapa* genome. **b** APA sites distribution mapped to *B.rapa* genome. **c** Novel isoforms density from the PacBio Sequel platform. **d** Novel loci density from the PacBio Sequel platform. The closer the color is to red, the higher the density. Conversely, the closer the color is to blue, the lower the density. **e** LncRNA density from the PacBio Sequel platform. The closer the point is to the center, the lower the density. **f** Fusion transcripts distribution. Purple line represents intra-chromosome fusion transcripts, and yellow line represents inter-chromosomal

were "general functional prediction only" (20.29%), "post-translational modification, protein turnover, chaperones" (16.71%), and "signal transduction mechanisms" (12.18%) (Fig. 7e).

LncRNA and ORF prediction of novel isoforms

LncRNAs have regulatory functions, and are crucial for post-transcription, transcription, and epigenetics [42]. The novel isoforms from novel genes and novel isoforms from known genes with no hit in the above functional annotation databases, were predicted by CPAT software to identify lncRNAs in the PacBio data set. To obtain a high confidence set of lncRNAs, we retained the isoforms with an optimum cutoff value, and that were more than 200 bp in length. A total of 407 novel isoforms were predicted to be lncRNAs, accounting for 1% of all novel isoforms, with a mean length of 1127 bp (Additional file 2: Table S8). About 168 lncRNAs (41.28%)

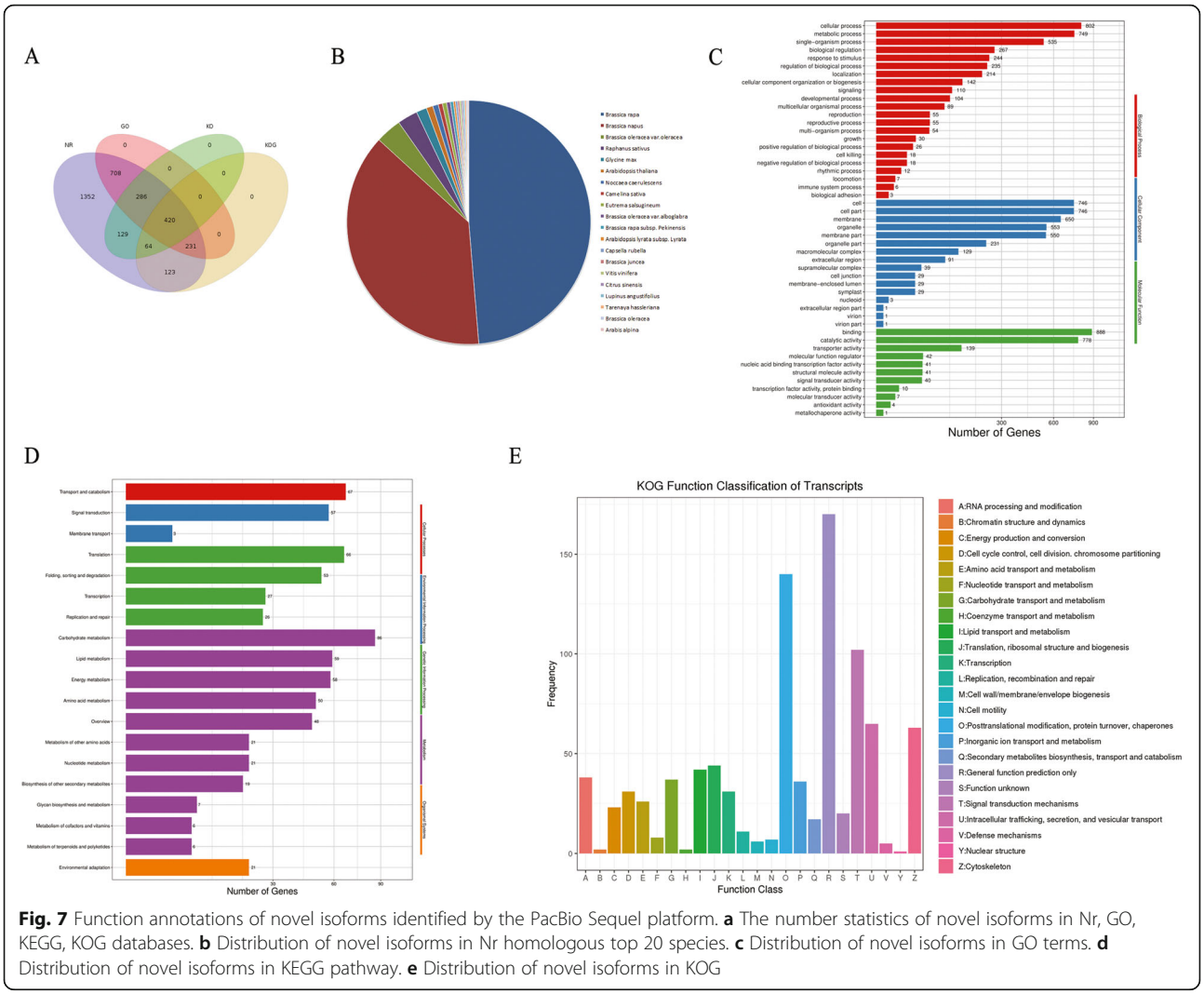

Fig. 7 Function annotations of novel isoforms identified by the PacBio Sequel platform. **a** The number statistics of novel isoforms in Nr, GO, KEGG, KOG databases. **b** Distribution of novel isoforms in Nr homologous top 20 species. **c** Distribution of novel isoforms in GO terms. **d** Distribution of novel isoforms in KEGG pathway. **e** Distribution of novel isoforms in KOG

were longer than 1000 bp, and four lncRNAs were longer than 4000 bp. The predicted lncRNAs were classified into four types, consisting of 33 antisense (8.11%), 289 intergenic (71.01%), 6 intronic (1.47%), and 79 sense (19.41%) lncRNAs (Fig. 8a). Mapping of the predicted lncRNAs to *B. rapa's* ten chromosomes was presented using Circos visualization software, revealing that 407 lncRNAs were randomly distributed, of which three lncRNAs were not anchored on the chromosomes (Fig.6e). The open reading frames (ORFs) were predicted by transDecoder software, resulting in 37,549 novel isoforms with a predicted ORF. Next, the density and length distributions of coding sequences (CDS) were investigated, and the mean length was 915 bp (Fig. 8b). The encoded peptide sequences are listed in Additional file 2: Table S9. The density and length of the distributions of the 5′ and 3′ boundaries of untranslated regions (UTRs) were identified, and the results revealed 415 3′ UTRs with a mean length of 641 bp and 8791 5′ UTRs with a mean length of 788 bp (Fig. 8c, d). Furthermore, the exon structures of novel isoforms with a predicted ORF and lncRNAs were analyzed, and the average number of exons per mRNA and lncRNA was 8.78 and 1.65, respectively (Fig. 8e).

Various models of AS

AS increases the diversity of transcriptomes and proteomes according to the different splice modes, rather than by massively amplifying the number of genes in cells or tissues [43, 44]. Traditionally, AS events consisted of several different types: exon skipping (SKIP) and cassette exons (MSIP), retention of single (IR) and multiple (MIR) introns, alternative exon ends (5′, 3′, or both) (AE), approximate exon skipping (XSKIP) and cassette exons (AMSKIP), approximate retention of single (XIR) and multiple (XMIR) introns, and approximate alternative exon ends (XAE) (Fig. 2c). The latest *B. rapa* v3.0 reference genome did not incorporate AS models

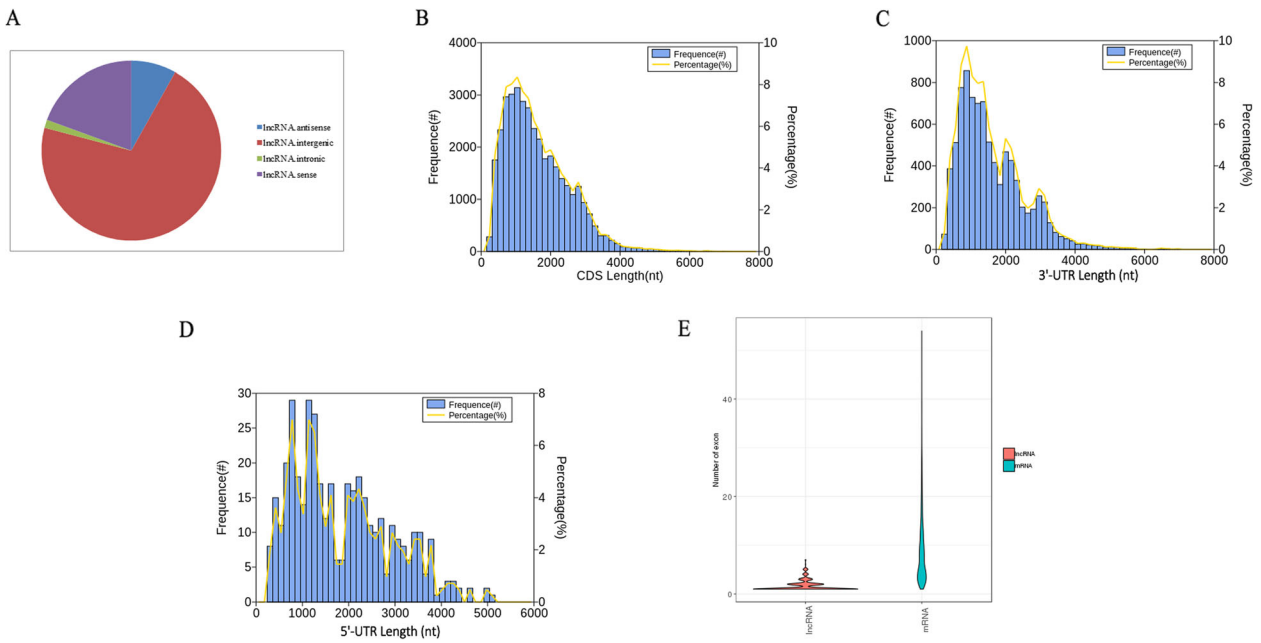

Fig. 8 LncRNA and ORF analysis. **a** Identification of four types of lncRNA. **b** Number, percentage and length distributions of CDS of novel isoforms with predicted ORF. **c** Number, percentage and length distributions of 3′ UTRs of novel isoforms with predicted ORF. **d** Number, percentage and length distributions of 5′ UTRs of novel isoforms with predicted ORF. **e** Exon number distribution of novel isoforms with predicted ORF and lncRNAs

and splice isoforms. However, there was a total of 156, 516 unique splice junctions detected in the earlier *B. rapa* genome v1.5, and IR events were predominant in the reference genome, similar to in species such as *M. truncatula*, *P. trichocarpa*, *A. thaliana*, *O. sativa*, *C. reinhardtii*, and *B. distachyon* [45]. In our study, we compared the PacBio-sequenced isoforms against the *B. rapa* genome v3.0, and found that a total of 19,503 loci corresponding to 51,501 isoforms underwent 453,270 AS events, indicating the distribution of AS events is much high in anthers (Additional file 2: Table S10). The total AS events generated in our study were: 1000 SKIP, 452 MSKIP, 15,592 IR, 4022 MIR, 5744 AE, 172,024 XKIP, 86,352 XMSKIP, 146,482 XIR, 2192 XMIR, and 19,410 XAE (Fig. 9a). Further, we observed that XSKIP (37.95%) was predominant, while MSKIP (0.1%) was the least frequent event. This findings greatly enriched anther transcript information. In our PacBio Sequel platform analysis, two or more isoforms were found in 3576 genes. Ten or more splice isoforms were detected in 1115 genes (Fig. 9b). The largest number of splice isoforms was 64,266, detected in *BraA06g022340.3C*; this gene is a homolog of *Arabidopsis* H(+)-ATPase 8 (AHA8). To verify the accuracy of AS events detected by SMRT, three genes were randomly selected, and gene-specific primers that spanned the predicted splicing events were designed for RT-PCR. The expression results for RT-PCR and Sanger sequencing in anthers were identical to the splice isoforms detected from PacBio data set, which demonstrated that those data were reliable (Additional file 1: Figure S1).

Fusion transcript and APA identification

A fusion transcript refers to a new gene formed by splicing together two or more separate genes, which were known as chimeric transcripts. The mechanisms leading to the generation of fusion transcripts include genomic structural variation, transposition, or trans-splicing after transcription. In this study, we identified 104 fusion transcripts, involving 187 annotated genes (Additional file 2: Table S11). Fusion transcripts were most frequently distributed on chromosome A03, followed by chromosome A09 and A01. According to the chromosomal distribution, we detected 101 inter-chromosome and 3 intra-chromosome fusion transcripts (Fig. 6f). This result was consistent with those of other species such as maize [14] and red clover [39]. Previous studies have indicated that most fusion transcripts are composed of two genes [46]. Consistent with these studies, all the 104 fusion transcripts in our data were composed of two genes. In addition, three fusion transcripts detected by SMRT were randomly selected and experimentally validated in anther and four other floral organs. The experimental results confirmed the authenticity of these chimeric RNAs (Additional file 1: Figure S1).

The post-transcriptional modification process of pre-mRNA to mature mRNA mainly includes the addition of a 7-methylguanosine cap at the 5′-end, intron splicing, and 3′-end formation by cleavage and polyadenylation [47].

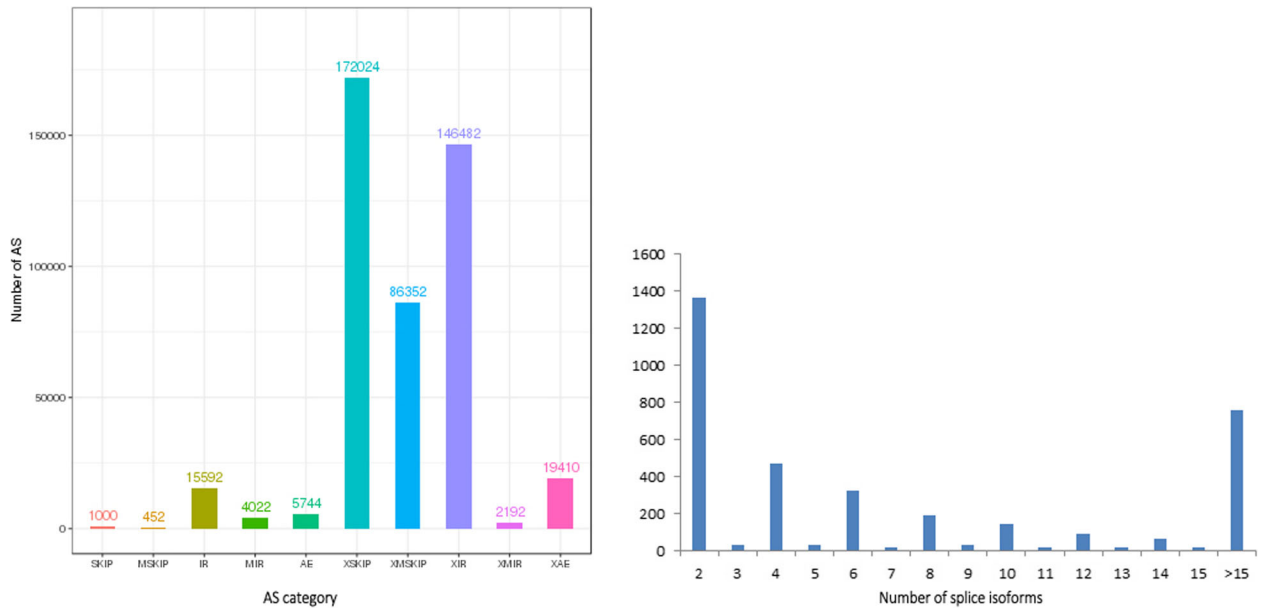

Fig. 9 Identification of AS events. **a** The number distribution of AS events in loci detected by the PacBio Sequel platform. **b** Distribution of loci that produce two or more splice isoforms detected by the PacBio Sequel platform

The specific position of the poly-A tail at the 3′-end is variable, and this variation may affect the binding of microRNA or RNA-binding protein to mRNA, and the process of RNA splicing and translation. The Tapis software was used to accurately identify polyadenylation sites in anthers. By investigating the 3′-end of transcripts in our PacBio data set, 24,816 poly-A sites were detected from 10,661 genes, of which 5806 genes have alternative polyadenylation (APA) (Fig. 6b, Additional file 2: Table S12). A total of 4855 genes have at least one poly-A site, while 733 genes have more than five poly-A sites (Fig. 10a). The mean number of poly-A sites per gene was 2.33. The largest number of poly-A sites was 19, found in *BraA02g029650.3C* and *BraA04g001890.3C*. We next analyzed the nucleotide distribution of the 50 nts in upstream and downstream of all poly-A sites. Consistent with results in other species, the poly-A sites from our PacBio data set showed a nucleotide bias, with an enrichment of uracil (U) upstream and adenine (A) downstream (Fig. 10b).

Discussion

At present, the reference genome of Chinese cabbage has been updated to version 3.0 using single-molecule sequencing. However, full-length transcripts, alterative spliced transcripts, fusion genes, and APA sites of Chinese cabbage have not been well-explored at the transcriptional level. Anthers are the male reproductive organs of plants that can produce pollen grains. The regulatory network of anther development is an extremely complex process involving a range of biological events [48]. In *Arabidopsis*, anther development has been divided into 14 stages, which make up two phases: microsporogenesis and microgametogenesis [49]. Briefly, anther development originates from stamen primordium formation, and microspore mother cells undergo meiosis to form haploid microspore tetrads. The microspores are surrounded by callose, and the release of individual microspores from the tetrads requires the action of callose enzyme secreted by the tapetum. Then, microspore wall is synthesized, followed by tapetum degradation, pollen mitosis divisions, septum cell degeneration, stomium differentiation, and finally anther dehiscence, releasing mature pollen grains. These events are relatively independent, and there is coordination in time and space. The abnormality of gene structure or expression in one of the events may cause loss of pollen function, which can generate male sterile lines. One crucial application of plant male sterility is hybrid seed production, and the advantage of hybrids is that they can increases seed yield and improve stress resistance [50, 51]. Therefore, it is necessary to investigate full-length mRNA information, providing a comprehensive view of splice isoforms in anther development.

PacBio sequencing is an effective platform for sequencing full-length transcripts because of its generation of long reads, which have an average length of 12 kb [52]. This long read length is why the PacBio sequencing platform can comprehensively analyze splice isoforms of each gene without assembly. In our work, we analyzed the full-length transcriptome of Chinese cabbage anther using the PacBio Sequel platform and yielded a total of 1,098,119 CCSs. Of these, 827,322 transcripts were identified as FLNCs, and the length of each sequencing

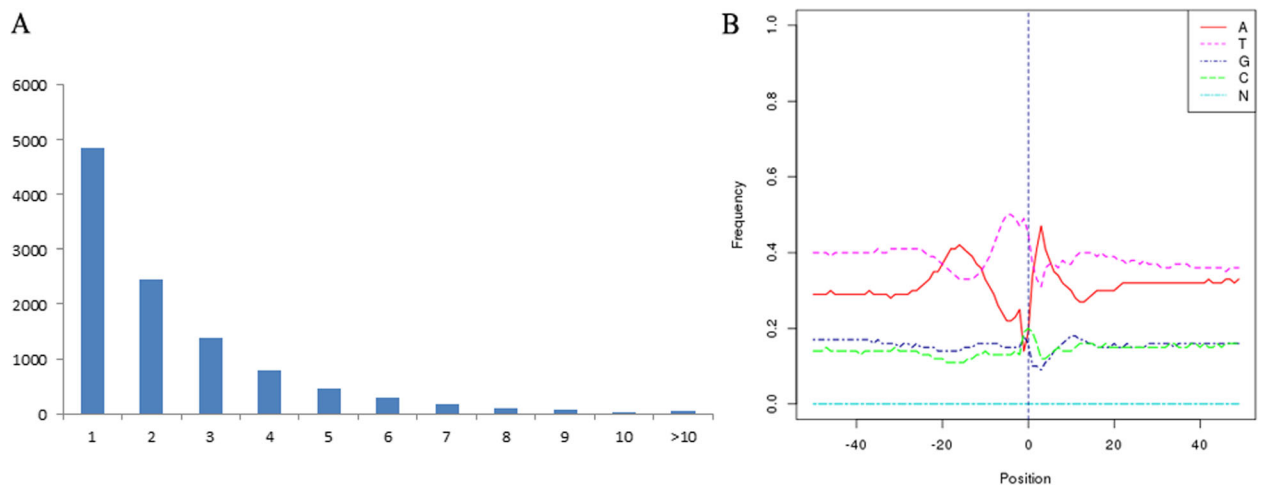

Fig. 10 APA analysis predicted by the PacBio Sequel platform. **a** The number distribution of poly-A sites per gene. **b** Nucleotide distribution around poly-A cleavage sites

library was consistent with the library standard (Fig. 3; Table 1). Single-molecule sequencing has a high base error rate of about 13%, mainly due to the addition of extra bases, especially in homopolymers [53]. However, as such errors occur randomly, there is no error bias, unlike that observed with NGS technology. Currently, the most common and effective method to further correct PacBio sequencing is to use high-accuracy data from an Illumina platform. With error correction using short-read RNA-Seq, 725,731 high quality FLNCs were identified to obtain 51,501 isoforms, consisting of 38,992 novel isoforms from 11,398 known genes and 3691 novel isoforms from 2682 novel genes (Additional file 2: Table S6). These results demonstrated that PacBio transcriptome sequencing can heighten the capacity to obtain full-length transcripts and enrich novel or uncharacterized isoforms or genes. Of the novel isoforms obtained, 407 high-confidence lncRNAs and 37,549 novel isoforms with predicted ORFs were identified (Additional file 2: Table S8 and Table S9). During pollen development and the fertilization process in *B. rapa*, a total of 12,501 putative lncRNAs were detected with an average length of 373 bp [42]. In our data, the mean length of predicted lncRNAs from novel isoforms was 1127 bp (Additional file 2: Table S8). Previously, the *B. rapa* genome was annotated using only ORFs, and thus there was no 5′ and 3′ UTRs defined. In 2013, Tong et al. provided a global transcriptional landscape in *B. rapa* accession Chiifu-401-42 and defined the 5′ and 3′ UTRs. The mean length of 5′ and 3′ UTRs was 139 bp and 184 bp, respectively [45]. In *Arabidopsis*, the mean length of 5′ and 3′ UTRs was 88 bp and 184 bp, respectively [54]. In our PacBio sequencing data, the mean length of 5′ and 3′ UTRs from novel isoforms with predicted ORF was 788 bp and 641 bp, respectively (Fig. 8c, d).

In addition to capturing full-length transcripts, another advantage of PacBio sequencing is the ability to detect AS events, which play a crucial role in regulating cellular molecules, cellular physiology, and developmental pathways [45, 55, 56]. The proportion of AS genes in rice, maize, *B. rapa*, and *A. thaliana* is 33, 37, 42, and 61%, respectively [57–59]. Limited by short reads, previous studies of the transcriptome using NGS technology have only been able provide individual splice junctions, while PacBio sequencing technology can be applied to alternatively spliced forms for each mRNA [39]. IR is the most common event in various genomes, which supports an intron-definition mechanism for pre-mRNA splicing [60]. In our study, we collected anthers from all developmental stages to harvest relatively comprehensive spliced isoforms. However, we detected a total of 453,270 AS events, and the majority of AS events were XSKIP (Fig. 9a). Previous studies indicated that alternative spliced transcripts have tissue-specific expression in various plants [61–64]. For novel splice junctions in *B. rapa*, 34.4% of alternative spiced transcripts were detected in only one tissue [45]. Therefore, differences in the prevalence of AS events may be related to tissue specificity. Those findings illustrate the complexity of the anther-specific transcriptome. Unfortunately, the expression levels of transcripts detected by PacBio sequencing have not been analyzed, and there is no way to analyze the expression pattern of different isoforms from one gene caused by AS events.

Taking the model plant *Arabidopsis* as an example, the key regulatory genes during anther development have been quite extensively reported, and are mainly involved in microsporogenesis, tapetum layer formation, callose layer development, pollen wall formation, and anther dehiscence [65]. Chinese cabbage and *Arabidopsis* both

belong to the Brassicaceae family, and so are closely related and have high sequence similarity. Therefore, we complied 34 genes that have been confirmed to be involved in anther development in *Arabidopsis* (Additional file 2: Table S13). In addition to the three whole genome duplications (WGDs) that occurred in Brassicaceae, the Brassica genome has undergone an additional ancient triplication, accompanied by gene fractionation [38]. Thus, based on the best BLASTX search in the Brassica database, we obtained 53 annotated genes from the PacBio annotation data (Additional file 2: Table S13). Of these genes, *AGAMOUS* (*AG*), *SPOROCYTELESS/NOZZLE* (*SPL*/*NZZ*), *BARELY ANY MERISTEM1/2* (*BAM1*/*2*), *Extra sporogenous cells*/*Excess microsporocytes 1* (*EMS1/EXS*), *SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE 1* (*SERK1*), and *TAPETUM DETERMINANT 1* (*TPD1*) were annotated for microsporogenesis in the early stages of anther development. For tapetal development and programmed cell death (PCD), the key genes detected were *Arabidopsis thaliana MYB DOMAIN PROTEIN 80/103* (*AtMYB80*/*AtMYB103*), *Dysfunctional Tapetum 1* (*DYT1*), *Tapetal Development and Functional 1* (*TDF1*), *Aborted microspores* (*AMS*), and *Male sterility* (*MS1*). For pollen exine formation, *Callose synthase 5* (*CALS5*), *Cyclin-dependent kinase G1* (*CDKG1*), *AUXIN RESPONSE FACTOR17* (*ARF17*), *No exine formation 1* (*NEF1*), *Ruptured pollen grains 1* (*RPG1*), *Defective in exine formation 1* (*DEX1*), *No primexine and plasma membrane undulation* (*NPU*), *CYP703A2, Acetyl-coenzyme A synthetase 5* (*ACOS5*), *MALE STERILITY 2* (*MS2*), *Less adherent pollen5* (*LAP5*), and *ATP-binding cassette G26* (*ABCG26/WBC27*) were detected. For pollen intine formation, *Cellulose synthase 1/3* (*CESA1*/*3*), *ARABINOGALACTAN PROTEIN 6/11* (*APG6*/*11*), and *Fasciclin-like arabinogalactan protein 3* (*FLA3*) were identified. For anther dehiscence, *MYB DOMAIN PROTEIN 26* (*MYB26*), and *NAC SECONDARY WALL THICKENING PROMOTING FACTOR 1* (*NST1*) were generated. Moreover, some of the loci were found to contain different alternatively spliced isoforms in our PacBio data set. For example, two loci (*BraA07g036270.3C* and *BraA07g029410.3C*) were annotated as *SERK1,* which is important for anther cell specification, but only *BraA07g036270.3C* expressed two alternatively spliced isoforms. As early as the meiosis phase, the callose layer begins to deposit outside the plasma membrane of microspore mother cells, which is the initiation of pollen wall development. In *Arabidopsis*, 12 *CALS* genes were identified, of which *CALS5* plays an important role in callose synthesis during the tetrad period. In mutant *cals5*, callose is insufficiently produced around microspores, resulting in defects in primexine formation and subsequently affecting the deposition of sporopollen in the pollen exine [66]. In our data, *CALS5* (*BraA09g010050.3C*) had ~ 1065 spliced variants, and XMSKIP predominated in AS models. For primexine formation, two loci (*BraA10g025410.3C* and *BraA02g004840.3C*) were annotated as *NEF1*. Two AS events, IR and XAE, were detected in *BraA10g025410.3C*, and XSKIP was found in *BraA02g004840.3C*. In *Arabidopsis*, multiple *CESA* genes encoding a cellulose synthase associated with pollen intine formation were cloned; the knockout mutant of *cesa1* and *cesa3* exhibited the gametophytic sterility phenotype, with abnormal pollen wall [67]. Both the annotated *CESA1* and *CESA3* in Chinese cabbage contained two loci each. *BraA01g005650.3C* was one of the *CESA1* loci, of which 40 alternatively spliced isoforms were detected, including twelve IR, six XIR, twelve XMSKIP, and ten XAE. *CESA1* (*BraA03g057280.3C)* had twelve alternatively spliced isoforms, consisting eight IR, two XIR, and two XAE. Similarly, two loci were annotated as *CESA3* (*BraA03g002020.3C* and *BraA02g001600.3C*). Eight AS events: six IR and two XAE were detected in *BraA03g002020.3C*, and two IR were detected in *BraA02g001600.3C*. Our research identified AS events in key genes active during anther development in Chinese cabbage.

Conclusions

Full-length transcriptome technology was used to explore the transcripts and splice isoforms present during anther development in Chinese cabbage. A total of 51,501 isoforms were identified using the PacBio Sequel platform. Meanwhile, 453,270 AS events were detected, and XSKIP events were found to have occurred extensively in anther. A total of 53 key genes active during anther development were detected in our PacBio sequencing, of which eight annotated loci had alternatively spliced isoforms. Additionally, 104 fusion transcripts and 24,816 poly-A sites were also predicted in this study. These new findings provide a valuable resource for complete characterization of anther-specific transcriptome data and improved Chinese cabbage genome annotation.

Methods

Plant material

The excellently Chinese cabbage DH line 'FT' was independently created by our laboratory (Liaoning Key Lab of Genetics and Breeding for Cruciferous Vegetable Crops) using isolated microspore culture technology. The DH line 'FT' is characterized by extremely early maturity, heat resistance, ovoid leaf head, and white petals (Fig. 1a). In August 2018, the DH line 'FT' seeds were placed in a 4 °C refrigerator for vernalization, and then sown in a greenhouse at Shenyang Agricultural University. At the full-bloom stage, three plants with consistent growth were randomly selected, and the whole buds of a complete inflorescence from each plant were individually collected in pieces of aluminum foil (Fig. 1b). Then, the anthers from each bud were detached, frozen in liquid

nitrogen, and stored at − 80 °C prior to SMRT sequencing (Fig. 1c).

PacBio library construction and sequencing

Total RNA from the three samples was extracted using Trizol reagent (Invitrogen, CA, USA). RNA purity and integrity was monitored by NanoPhotometer® spectrophotometer (IMPLEN, CA, USA) and a Bioanalyzer 2100 system (Agilent Technologies, CA, USA). RNA contamination was assessed by 1% agarose gel. RNA concentration was detected using a Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). Equimolar rations of the total RNA from each sample were mixed together. The full-length cDNA was prepared using a SMARTer™ PCR cDNA Synthesis Kit (Takara Biotechnology, Dalian, China). Size fractionation (1–2, 2–3, and > 3) of full-length cDNA was achieved using the BluePippin™ Size Selection System (Sage Science, Beverly, MA). The filtered full-length cDNAs were subjected to re-amplification, end repair, SMRT adapter ligation, and exonuclease digestion. After secondary screening by BluePippin™, three SMRTbell libraries were constructed with the Pacific Biosicences DNA Template Prep Kit 2.0. Library quantification and size was measured using a Qubit® 2.0 Fluorometer (Life Technologies, CA, USA) and Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Subsequently, SMRT sequencing was performed on a PacBio Sequel platform by Frasergen Bioinformatics Co., Ltd. (Wuhan, China).

Illumina RNA-Seq library construction and sequencing

In parallel, the quantity and purity of equally mixed RNA were analyzed using Bioanalyzer 2100 and RNA 6000 Nano LabChio Kit (Agilent, CA, USA). Poly (A) mRNA was isolated by poly-T oligoattached magnetic beads (Invitrogen). Following fragmentation, the cleaved RNA fragments were reverse-transcribed into a cDNA library following treatment with the mRNASeqample Preparation Kit (Illumina, San Diego, USA). After assessing the library quality, we performed PE300 sequencing on an Illumina Hiseq 2500 at the LC Sceiences (Hangzhou, China) following the vendor's recommended protocol.

Quality filtering and error correction

PacBio raw reads were preprocessed and filtered using SMRT Link v5.0. Briefly, CCSs were generated from the subread SAM file with the following parameters: minimum subread length = 50; minimum number of passes = 1, minimum predicted accuracy = 0.8, and minimal read score = 0.65. Then, CCSs were classified into either full length or non-full length reads, by assessing the presence of the 5′ and 3′ adapters and poly (A) tail. FLNC reads were full length CCSs containing all three elements, with no additional copies of the adapter sequence within the DNA fragment.

The high-quality Illumina short reads were used to error correction for FLNC reads. The proovread software v2.12 was widely and efficiently applied for correcting FLNC sequences by iterative short read consensus [68]. Using GMAP[2] software, the FLNC sequences before and after error correction were compared to the *B. rapa* v3.0 reference genome [69, 70] with "—no-chimeras and –n 100" to calculate the PID values, including global PID and local PID (Fig. 2b). The higher the PID values, the more consistent the sequencing data was with the reference genome. The PID values of the genomic comparison before and after error correction were separately counted, and the higher PID values were updated. Then, the uniquely mapped FLNC sequences with high PID (global PID > 95% and local PID > 97%) were used to annotate loci and isoforms.

Gene loci and isoform finding

Gene loci and isoforms were identified based on the alignment position of the corrected FLNC reads. For loci, two transcripts that overlapped by at least 20% of their initiation sites on the same strand, and had at least one exon overlap of more than 20%, were considered to be the same loci transcript. These same loci transcripts were further analyzed for isoform identification. The process mainly included the removal of redundant transcripts and the filtration of low-reliability transcripts. The redundant transcripts were removed as follows: firstly, if all the splicing sites of the same loci transcripts were identical, they could be considered one isoform; secondly, if one isoform was degraded at the 5′ terminal region, but the remaining region was consistent with other isoforms, it should be filtered out. For false positives, when the global PID < 99%, each isoform structural model must supported with least two FLNC reads; otherwise, if there was only one sequence, then all junction sits of the sequence were fully supported by the genomic annotation or Illumina RNA-Seq data.

Novel gene and isoform identification

The above gene loci and isoforms were compared with the reference annotation to identify known genes and isoforms, as well as novel genes and isoforms. A sequenced gene was determined to be a novel gene by satisfying any of the following criteria: (i) There is no overlap or there is an overlap of less than 20% of the annotated genes; or (ii) the overlap with the annotated gene is greater than 20%, but the gene direction is inconsistent. In addition, if the sequenced isoform contained one or more new splicing sites, or if the sequenced isoform and annotated isoform were not both single-exon, it was considered to be a novel isoform.

Functional annotation

The novel isoforms were annotated by NR, KOG, KO, and Swiss-Prot databases with Diamond [71, 72]. KEGG pathways were searched by KOBAS v2.0 [73]. GO annotations were performed by BLASTX v2.2.26 and BLAST2GO v2.3.5 software [74].

LncRNA and ORF identification

To identify LncRNA, novel isoforms from known genes or novel isoforms from novel genes obtained by PacBio data were first searched against NR, KOG, KO, and Swiss-Prot databases with default parameters. The isoforms that had BLAST hits with 1E-5 were filtered out, and the remaining isoforms were further evaluated for protein-coding capacity by CPAT v1.2.2 (http://lilab.research.bcm.edu/cpat/).

To predict ORFs, transDecoder software was used to identify potential coding sequences (http://transdecoder.sf.net). By default, the length of ORFs predicted by TransDecoder.LongOrfs was at least 100 amino acids. To improve the sensitivity of ORFs, possible ORF-translated proteins were aligned to the Swiss-Prot database by BlastP for homologous protein identification. Simultaneously, protein domain identification was determined from the Pfam database by Hmmscan [75, 76]. Subsequently, TransDecoder. Predict was used to filter all predicted ORFs based on the above results, and retained ORFs that have homology to the Swiss-Prot database or with the same domain.

AS, fusion transcript, and APA identification

Alternative splicing (AS) events were ascertained using ASprofile software [77]. The splice types, (M) SKIP, (M) IR, AE, X (M) KIP, X (M) IR, and XAE were classified and characterized by comparing different isoforms at the same gene loci using ASprofile (Fig. 2c). Fusion transcripts were those where the 5′ and 3′ sequences mapped to two or more gene loci in the reference genome, corresponding to the 5′ partner and 3′ partner genes. The iso-seq fusion transcripts detection software, self-developed by Frasergen Inc. (Wuhan, China), was used for fusion gene detection. A schematic diagram of the software is shown in Fig. 2d. Poly-A site was an important post-transcriptional modification of RNA. The reliable APA sites were obtained by Tapis software [33].

RT-PCR validation

Total RNA from floral organs of the DH line 'FT', including anthers, sepal, filament, petal, and pistil were extracted and mixed as described above. Reverse transcription was conducted using the FastQuant RT Super Mix (TIANGEN, China). RT-PCR was performed in 10 μl volumes containing 50 ng DNA, 1.0 μl of 10 Taq Reaction Buffer (containing $Mg2^{+}$), 0.8 μl of 2.5 mM dNTP, 1 μl each of 0.5 μm forward and reverse primers, and 1 U of Taq DNA polymerase (TIANGEN, China). The amplification was performed on an iCycler thermocycler (Bio-Rad, USA) with the following cycling parameters: initial denaturation at 95 °C for 5 min, and 35 cycles of 95 °C for 30 s, 56 °C for 30 s, and 72 °C for 30 s, with a final extension at 72 °C for 10 min. Gene-specific primers were designed with Primer Premier 5.0 by GENEWIZ (Suzhou, China). PCR products were analyzed on 2% agarose gels and followed by Sanger sequencing. All primers are listed in Additional file 2: Table S14.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12870-019-2133-z.

Additional file 1 Figure S1. RT-PCR validation of AS events (1–3) and fusion transcripts (4–6). M, DNA Marker DL2000; A, anther; S, sepal; F, filament; Pe, petal; Pi, pistil; 1, m54191_180531_084316/71238183/3459_97_CCS; 2, m54191_180531_084316/15467311/43_3034_CCS; 3, m54045_180508_172253/21365668/2097_84_CCS; 4, *BraA03g009520.3C*; 5, *BraA01g012300.3C*; 6, *BraA02g020980.3C*.

Additional file 2 Table S1. Summary of polymerase reads from the PacBio Sequel platform. **Table S2.** Summary of subreads from the PacBio Sequel platform. **Table S3.** Summary of CCSs from the PacBio Sequel platform. **Table S4.** Global PID statistics before and after sequencing error correction. **Table S5.** Evaluation of full-length transcripts in the PacBio data set. **Table S6.** Classification of loci and isoforms mapped to the reference genome. **Table S7.** Functional annotation of all novel isoforms by the PacBio Sequel platform. **Table S8.** Information of lncRNAs from the PacBio Sequel platform. **Table S9.** ORF information detected by the PacBio Sequel platform. **Table S10.** Splice isoforms detected by PacBio Sequel platform. **Table S11.** Fusion genes detected by the PacBio Sequel platform. **Table S12.** Poly-A sites detected by the PacBio Sequel platform. **Table S13**. Anther and pollen development related genes in *B. rapa* genome v3.0. **Table S14**. Primers used for RT-PCR validation.

Abbreviations

APA: alternative polyadenylation; AS: alternative splicing; CCS: circular consensus sequence; FLNC: full-length non-chimeric; LncRNA: long non-coding RNA; NGS: next-generation sequencing; ORF: open reading frame; PID: percentage-of-identity; SMRT: single-molecule real-time sequencing

Acknowledgements

We acknowledge Yanbo Feng from Frasergen Bioinformatics Co., Ltd. (Wuhan, China) for providing relevant literature regarding the PacBio sequel platform, and actively coordinating communication with technical staff to facilitate the completion of this manuscript. We also thank Editage (www.editage.cn) for English language editing.

Authors' contributions

ZL and FH conceived and designed this study. CT analyzed the data and wrote the manuscript. CT and HL performed the verification experiments. CT, JR, and XY performed data analysis. All the authors read and approved the final manuscript.

Funding

This work was supported by the National Key Research and Development Program of Chinese (No. 2016YFD0101701) and National Natural Science Foundation of China (No. 31772298 and No. 31201625). Each of the funding bodies granted the funds based on a research proposal. They had no influence over the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. We deposited the raw SMRT data in the Sequence Read Archives (SRA) of the National Center for Biotechnology Information (NCBI) under the accession numbers SRR10259626, SRR10259627 and SRR10259628 of Bioproject PRJNA576779. The Illumina RNA-Seq data was uploaded to the SRA under the accession number SRR10247439 of the Bioproject ID PRJNA576332. Genomic sequences and gene annotation information of *B.rapa* are downloaded online at http://brassicadb.org/brad/datasets/pub/Genomes/Brassica_rapa/V3.0/.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All the authors declare that they have no competing interests.

Received: 27 July 2019 Accepted: 12 November 2019
Published online: 27 November 2019

References

1. Sang F, Nicklen S. Coulson. DNA sequencing with chain-terminating inhibitors. PNAS. 1977;74(12):5463–7.
2. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next generation sequencing technology. Trends Genet. 2014;30(9):418–26.
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;15(1):57–63.
4. An H, Yang Z, Yi B, Wen J, Shen J, Tu J, Ma C, Fu T. Comparative transcript profiling of the fertile and sterile flower buds of pol CMS in *B napus*. BMC Genomics. 2014;15:258.
5. Liu C, Liu Z, Li C, Zhang Y, Feng H. Comparative transcriptome analysis of fertile and sterile buds from a genetically male sterile line of Chinese cabbage. In Vitro Cell Dev Biol Plant. 2016;52(2):130–9.
6. Liu XQ, Yu CY, Dong JG, Xu AX, Hu SW. *De novo* transcriptome reconstruction of a thermo-sensitive male sterility mutant in rapeseed (*Brassica napus*; Brassicaceae). App Plant Sci. 2017;5(12).pii:apps.1700077.
7. Pei X, Jing Z, Tang Z, Zhu Y. Comparative transcriptome analysis provides insight into differentially expressed genes related to cytoplasmic male sterility in broccoli (*Brassica oleracea* var. italica). Sci Horic. 2017;217:234–42.
8. Wang S, Wang C, Zhang XX, Chen X, Liu JJ, Jia XF, Jia SQ. Transcriptome *de novo* assembly and analysis of differentially expressed genes related to cytoplasmic male sterility in cabbage. Plant Physiol Bioch. 2016;105:224–32.
9. Xu HM, Kong XD, Chen F, Huang JX, Lou XY, Zhao JY. Transcriptome analysis of *Brassica napus* pod using RNA-Seq and identification of lipidrelated candidate genes. BMC Genomics. 2015;16(1):1–10.
10. Yan X, Dong C, Yu J, Liu W, Jiang C, Liu J, Hu Q, Fang X, Wei W. Transcriptome profile analysis of young floral buds of fertile and sterile plants from the self-pollinated offspring of the hybrid between novel restorer line NR1 and Nsa CMS line in *Brassica napus*. BMC Biochem. 2013;14(3):1–16.
11. Zhou X, Liu Z, Ji R, Feng H. Comparative transcript profiling of fertile and sterile flower buds from multiple-allele-inherited male sterility in Chinese cabbage (*Brassica campestris* L. ssp. *pekinensis*). Mol Gen Genomics. 2017;292(5):967–90.
12. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
13. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31(11):1009.
14. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 2016;7:11708.
15. Dong L, Liu H, Zhang J, Yang S, Kong G, Chu JS, Chen N, Wang D. Single molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. BMC Genomics. 2015;16:1039.
16. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN. Evaluation of *de nove* transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15(12):553.
17. Zhu FY, Chen MX, Ye NH, Shi L, Ma KL, Yang JF, Cao YY, Zhang YJ, Yoshida T, Fernie AR, Fan GY, Wen B, Zhou R, Liu TY, Fan T, Gao B, Zhang D, Hao GF, Xiao S, Liu YG, Zhang J. Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings. Plant J. 2017;91(3):518–33.
18. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227–40.
19. Pushkarev D, Neff NF, Quake DR. Single-molecule sequencing of an individual human genome. Nat Biotechnol. 2009;27(9):847–50.
20. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biol. 2013;14(7):405.
21. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018;34(9):666–81.
22. Allen SL, Delaney EK, Kopp A, Chenoweth SF. Single-Molecule Sequencing of the Drosophila serrata Genome. G3 (Bethesda). 2017;7(3):781–788.
23. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J, Barker T, Lu FH, McKenzie N2, Raats D, Ramirez-Gonzalez RH, Coince A, Peel N, Percival-Alwyn L, Duncan O, Trösch J3, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Palma FD, Phillips AL, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW, Clark MD. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res. 2017;27(5):885–96.
24. Csabai Z, Tombácz D, Deim Z, Snyder M, Boldogkői Z. Analysis of the complete genome sequence of a novel. Pseudorabies Virus Strain Isolated in Southeast Europe Can J Infect Dis Med Microbiol. 2019; 2019:1806842.
25. Edger PP, VanBuren R, Colle M, Poorten TJ, Wai CM, Niederhuth CE, Alger EI, Ou S, Acharya CB, Wang J, Callow P, McKain MR, Shi J, Collier C, Xiong Z, Mower JP, Slovin JP, Hytönen T, Jiang N, Childs KL, Knapp SJ. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. Gigascience. 2018;7(2):1–7.
26. Li Y, Wei W, Feng J, Luo H, Pi M, Liu Z, Kang C. Genome re-annotation of the wild strawberry Fragaria vesca using extensive Illumina-and SMRT-based RNA-seq datasets. DNA Res. 2017. https://doi.org/10.1093/dnares/dsx038.
27. Peng Z, Hu Y, Xie J, Potnis N, Akhunova A, Jones J, Liu Z, White FF, Liu S. Long read and single molecule DNA sequencing simplifies genome assembly and TAL effectorgene analysis of *Xanthomonas translucens*. BMC Genomics. 2016;17:21.
28. Prakash G, Kumar A, Sheoran N, Aggarwal R, Satyavathi CT, Chikara SK, Ghosh A, Jain RK. First Draft Genome Sequence of a Pearl Millet Blast Pathogen, Magnaporthe grisea Strain PMg_DI, Obtained Using PacBio Single-Molecule Real-Time and Illumina NextSeq 500 Sequencing. Microbiol Resour Announc. 2019;8(20). pii: e01499–18.
29. Zhang L, Hu J, Han X, Li J, Gao Y, Richards CM, Zhang C, Tian Y, Liu G, Gul H, Wang D, Tian Y, Yang C, Meng M, Yuan G, Kang G, Wu Y, Wang K, Zhang H, Wang D, Cong P. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nat Commun. 2019;10(1):1494.
30. Zhao L, Deng L, Li G, Jin H, Cai J, Shang H, Li Y, Wu H, Xu W, Zeng L, Zhang R, Zhao H, Wu P, Zhou Z, Zheng J, Ezanno P, Yang AX, Yan Q, Deem MW, He J. Single molecule sequencing of the M13 virus genome without amplification. PLoS One. 2017;12(12):e0188181.
31. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M. Quantification of the yeast transcriptome by single-molecule sequencing. Nat Biotechnol. 2009;27(7):652–8.
32. Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, Zhu Y, Ji A, Zhang B, Hu S, Au KF, Song J, Chen S. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis. Plant J. 2015;82(6):951–61.
33. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, BenHur A, Reddy AS. A survey of the sorghum transcriptome using singlemolecule long reads. Nat Commun. 2016;7:11706.
34. Wang T, Wang H, Cai D, Gao Y, Zhang H, Wang Y, Lin C, Ma L, Gu L. Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (Phyllostachys edulis). Plant J. 2017;91(4):684–99.
35. He L, Fu S, Xu Z, Yan J, Xu J, Zhou H, Zhou J, Chen X, Li Y, Au KF, Yao H. Hybrid Sequencing of Full-length cDNA Transcripts of Stems and Leaves in Dendrobium officinale. Genes (Basel). 2017;8(10). pii: E257.

36. Chao Y, Yuan J, Li S, Jia S, Han L, Xu L. Analysis of transcripts and splice isoforms in red clover (*Trifolium pretense* L.) by single-molecule long-read sequencing. BMC Plant Biol. 2018;18(1):300.
37. Chao Y, Yuan J, Guo T, Xu L, Mu Z, Han L. Analysis of transcripts and splice isoforms in *Medicago sativa* l.by single-molecule long-read sequencing. Plant Mol Biol. 2019;99(3):219–35.
38. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang H, Wang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Wang H, Jin H, Parkin IA, Batley J, Kim JS, Just J, Li J, Xu J, Deng J, Kim JA, Li J, Yu J, Meng J, Wang J, Min J, Poulain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiary N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Zhang S, Huang S, Sato S, Sun S, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y, Qiu Y, Yin Y, Li Y, Du Y, Liao Y, Lim Y, Narusaka Y, Wang Y, Wang Z, Li Z, Wang Z, Xiong Z, Zhang Z; *Brassica rapa* Genome Sequencing Project Consortium. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet. 2011;43(10):1035–9.
39. Cai C, Wang X, Liu B, Wu J, Liang J, Cui Y, Cheng F, Wang X. *Brassica rapa* 2.0: a reference upgrade through sequence re-assembly and gene re-annotation. Mol Plant. 2017;10(4):649–51.
40. Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, Liang J, Cai C, Liu Z, Liu B, Wang F, Li S, Liu F, Li X, Cheng L, Yang W, Li MH, Grossniklaus U, Zheng H, Wang X. Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosomeconformation capture technologies. Hortic Res. 2018;5:50.
41. Ning G, Cheng X, Luo P, Liang F, Wang Z, Yu G, Li X, Wang D, Bao M. Hybrid sequencing and map finding (HySeMaFi): optional strategies for extensively decipheringgene splicing and expression in organisms without reference genome. Sci Rep. 2017;7:43793.
42. Huang L, Dong H, Zhou D, Li M, Liu Y, Zhang F, Feng Y, Yu D, Lin S, Cao J. Systematic identification of long non-coding RNAs during pollen development and fertilization in *Brassica rapa*. Plant J. 2018;96(1):203–22.
43. Kelemen O, Convertini P, Zhang ZY, Wen Y, Shen ML, Falaleeva M, Stefan S. Function of alternative splicing. Gene. 2013;514(1):1–30.
44. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470–6.
45. Tong C, Wang X, Yu J, Wu J, Li W, Huang J, Dong C, Hua W, Liu S. Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in *Brassica rapa*. BMC Genomics. 2013;14:689.
46. Weirather JL, Afshar PT, Clark TA, Tseng E, Powers LS, Underwood JG, Zabner J, Korlach J, Wong WH, Au KF. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. Nucleic Acids Res. 2015;43(18):e116.
47. de Almeida SF, García-Sacristán A, Custódio N, Carmo-Fonseca M. A link between nuclear RNA surveillance, the human exosome and RNA polymerase II transcriptional termination. Nucleic Acids Res. 2010;38(22):8015–26.
48. Scott RJ, Spielman M, Dickinson HG. Stamen structure and function. Plant Cell. 2004;16(suppl):S46–60.
49. Sander.
50. Schnable PS, Spriger NM. Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol. 2013;64:71–88.
51. Verma N. Transcriptional regulation of anther development in *Arabidpsis*. Gene. 2019;689:202–9.
52. Vembar SS, Seetin M, Lambert C, Nattestad M, Schatz MC, Baybayan P, Scherf A, Smith ML. Complete telomere-to-relomere *de novo* assembly of the *Plasmodium falciparum* genome through long-read (>11kb), single molecule, real-time sequencing. DNA Res. 2016;23(4):339–51.
53. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of ion torrent. BMC Genomics. 2012;13:341.
54. Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA. Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. Plant Mol Biol. 2006;60(1):69–85.
55. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet. 2011;12(10):715–29.
56. Seo PJ, Park MJ, Park CM. Alternative splicing of transcription factors in plant responses to low temperature stress: mechanisms and functions. Planta. 2013;237(6):1415–24.
57. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res. 2010;20(1):45–58.
58. Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR, Reidel EJ, Turgeon R, Liu P, Sun Q, Nelson T, Brutnell TP. The developmental dynamics of the maize leaf transcriptome. Nat Genet. 2010;42(12):1060–7.
59. Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L, Tian W, Tao Y, Kristiansen K, Zhang X, Li S, Yang H, Wang J, Wang J. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res. 2010;20(5):646–54.
60. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. Genome-wide analysis of transcript isoform variation in humans. Nat Genet. 2008;40(2):225–31.
61. Chao Q, Gao ZF, Zhang D, Zhao BG, Dong FQ, Fu CX, Liu LJ, Wang BC. The developmental dynamics of the Populus stem transcriptome. Plant Biotechnol J. 2019;17(1):206–19.
62. Qiao D, Yang C, Chen J, Guo Y, Li Y, Niu S, Cao K, Chen Z. Comprehensive identification of the full-length transcripts and alternative splicing related to the secondary metabolism pathways in the tea plant (Camellia sinensis). Sci Rep. 2019;9(1):2709.
63. Sun Y, Hou H, Song H, Lin K, Zhang Z, Hu J, Pang E. The comparison of alternative splicing among the multiple tissues in cucumber. BMC Plant Biol. 2018;18(1):5.
64. Wang M, Wang P, Liang F, Ye Z, Li J, Shen C, Pei L, Wang F, Hu J, Tu L, Lindsey K, He D, Zhang X. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. New Phytol. 2018;217(1):163–78.
65. Wilson ZA, Song J, Taylor B, Yang C. The final split: the regulation of anther dehiscence. J Exp Bot. 2011;62:1633–49.
66. Dong X, Hong Z, Sivaramakrishnan M, Mahfouz M, Verma DP. Callose synthase (*CalS5*) is required for exine formation during microgametogenesis and for pollen viability in *Arabidopsis*. Plant J. 2005;42(3):315–28.
67. Persson S, Paredez A, Carroll A, Palsdottir H, Doblin M, Piondexter P, Khitrov N, Auer M, Somerville CR. Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in *Arabidopsis*. PNAS. 2007;104(39):15566–71.
68. Hackl T, Hedrich R, Schultz J, Forster F. Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics. 2014;30(21):3004–11.
69. Wu TD, Wantanabe CK. GMAP: a genome mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.
70. Yuan D, Tang Z, Wang M, Gao W, Tu L, Jin X, Chen L, He Y, Zhang L, Zhu L, Li Y, Liang Q, Lin Z, Yang X, Liu N, Jin S, Lei Y, Ding Y, Li G, Ruan X, Ruan Y, Zhang X. The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. Sci Rep. 2015;5:17662.
71. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methoed. 2015;12(1):59–60.
72. Gasteiger E, Jung E, Bairoch A. SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol. 2001;3(3):47–55.
73. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L , Gao G, Li CY, Wei L. KOSAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011;39(Web Server issue): W316–22.
74. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–6.
75. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23(1):205–11.
76. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res Database Issue. 2010;38:D211–22.
77. Florea L. Song L and Salzberg SL. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. F1000Res. 2013; 2:188.

Publisher's Note