Single-molecule real-time sequencing identifies massive full-length cDNAs and alternative-splicing events that facilitate comparative and functional genomics study in the hexaploid crop sweet potato

Na Ding[1,2], Huihui Cui[1], Ying Miao[1], Jun Tang[3,4], Qinghe Cao[3,4] and Yonghai Luo[1,2]

[1] Fujian Provincial Key Laboratory of Plant Functional Biology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China

[2] School of Life Sciences, Jiangsu Normal University, Xuzhou, Jiangsu, China

[3] Jiangsu Xuhuai Regional Xuzhou Institute of Agricultural Sciences, Xuzhou, Jiangsu, China

[4] Key Laboratory of Biology and Genetic Improvement of Sweetpotato, Ministry of Agriculture, Xuzhou, Jiangsu, China

ABSTRACT

Background. Sweet potato (*Ipomoea batatas* (L.) Lam.) is one of the most important crops in many developing countries and provides a candidate source of bioenergy. However, neither a complete reference genome nor large-scale full-length cDNA sequences for this outcrossing hexaploid crop are available, which in turn impedes progress in research studies in *I. batatas* functional genomics and molecular breeding.

Methods. In this study, we sequenced full-length transcriptomes in *I. batatas* and its diploid ancestor *I. trifida* by single-molecule real-time sequencing and Illumina second-generation sequencing technologies. With the generated datasets, we conducted comprehensive intraspecific and interspecific sequence analyses and experimental characterization.

Results. A total of 53,861/51,184 high-quality long-read transcripts were obtained, which covered about 10,439/10,452 loci in the *I. batatas*/*I. trifida* genome. These datasets enabled us to predict open reading frames successfully in 96.83%/96.82% of transcripts and identify 34,963/33,637 full-length cDNA sequences, 1,401/1,457 transcription factors, 25,315/27,090 simple sequence repeats, 1,656/1,389 long non-coding RNAs, and 5,251/8,901 alternative splicing events. Approximately, 32.34%/38.54% of transcripts and 46.22%/51.18% multi-exon transcripts underwent alternative splicing in *I. batatas*/*I. trifida*. Moreover, we validated one alternative splicing event in each of 10 genes and identified tuberous-root-specific expressed isoforms from a starch-branching enzyme, an alpha-glucan phosphorylase, a neutral invertase, and several ABC transporters. Overall, the collection and analysis of large-scale long-read transcripts generated in this study will serve as a valuable resource for the *I. batatas* research community, which may accelerate the progress in its structural, functional, and comparative genomics studies.

Submitted 14 May 2019
Accepted 23 September 2019
Published 15 November 2019

Corresponding author
Yonghai Luo, yonghailuo@foxmail.com

Academic editor
Yuriy Orlov

Additional Information and Declarations can be found on page 12

DOI 10.7717/peerj.7933

Copyright
2019 Ding et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

How to cite this article Ding N, Cui H, Miao Y, Tang J, Cao Q, Luo Y. 2019. Single-molecule real-time sequencing identifies massive full-length cDNAs and alternative-splicing events that facilitate comparative and functional genomics study in the hexaploid crop sweet potato. *PeerJ* 7:e7933 http://doi.org/10.7717/peerj.7933

Subjects Agricultural Science, Genomics, Plant Science
Keywords Single-molecular real-time sequencing, Comparative transcriptome analysis, Alternative splicing, Sweet potato

INTRODUCTION

Sweet potato (*Ipomoea batatas* (L.) Lam.) is the seventh most important crop in the world and it ensures food supply and safety in many developing countries. *I. batatas* is a hexaploid plant with a complex and heterozygous genome (2n = 6 × = 90, 3–4 gigabase pairs in genome size (*Magoon, Krishnan & Vijaya, 1970*; *Ozias-Akins & Jarret, 1994*)). A preliminary genome estimate has revealed two genome polyploidization events occurring about 0.8 and 0.5 million years ago (*Yang et al., 2017*). Nevertheless, the complete reference genome of *I. batatas* remains lacking, which hinders the progress in molecular dissections of its evolutionary scenario and agronomically important traits. Moreover, *I. batatas* is a self-incompatible and thus obligate, outcrossing species (*Martin, 1965*). It is almost impossible to develop typical mapping populations such as F2 and recombinant inbred lines for constructing high-density linkage maps and classical genetic analyses. To date, no successful investigation in forward genetics (i.e., quantitative trait locus mapping and subsequently map-based cloning) of *I. batatas* has been reported. Therefore, RNA sequencing (i.e., RNA-seq, whole transcriptome shotgun sequencing (*Wang, Gerstein & Snyder, 2009*)) has been widely used as an attractive alternative to whole genome sequencing for gene mining in *I. batatas* (*Schafleitner et al., 2010*; *Wang et al., 2010*; *Nurit et al., 2013*). However, all reported transcriptomes in *I. batatas* were derived from second-generation sequencing platforms, which generate relatively short reads (i.e., hundreds of base pairs per read) and are disadvantageous in obtaining full-length transcripts (*Koren et al., 2012*). To date, the collection and analysis of large-scale full-length cDNA sequences have not been done in *I. batatas*, which is fundamental to its structural and functional genomics studies.

Ipomoea trifida (H.B.K.) G. Don has been considered as the diploid ancestor of *I. batatas* and accumulative evidence supports this hypothesis (*Srisuwan, Sihachakr & Siljak-Yakovlev, 2006*; *Wu et al., 2018*). Nevertheless, the evolutionary scenario underlying the origin and domestication of *I. batatas* remains unclear. Unlike *I. batatas*, *I. trifida* does not form tuberous roots, and thus comparative analysis of *I. batatas* and *I. trifida* may provide insights into the evolution and domestication of *I. batatas*. Although the reference genome of *I. trifida* becomes available recently (*Wu et al., 2018*) and short-read transcriptomes of *I. trifida* have been analyzed in a few projects (*Cao et al., 2016*; *Ponniah et al., 2017*), no study involving the large-scale collection and analysis of full-length cDNA sequences in *I. trifida* has been reported.

Long-read or full-length cDNA sequences are fundamental to structural and functional genomics studies. First, they provide complete information of transcribed sequences, which are required to gene function analyses. Second, they facilitate accurate predictions of gene models (i.e., to define proper orientation, order, and boundary of exons). Third, they may be utilized in validating or correcting the scaffold assembly in genome sequencing projects. Fourth, they are particularly useful to analyze alternative splicing of transcript isoforms, which is important to increase transcriptome diversity and adaptation potential of an organism. In the past, collecting full-length cDNA sequences was expensive, labor intensive, and time consuming (*Seki et al., 2002*; *Shoshi et al., 2003*). The advent of a third-generation sequencing platform (i.e., single-molecule real-time (SMRT) sequencing) has revolutionized DNA sequencing and thus genome/transcriptome studies (*Eid et al., 2010*). Long reads of up to 20-kb in size, albeit with a relatively high error rate, can be produced by SMRT sequencing (*Roberts, Carneiro & Schatz, 2013*; *Au et al., 2013*). Today, high-throughput sequencing combining second-generation sequencing (to generate short reads with high base quality) and SMRT sequencing (to produce long reads with a relatively high error rate) has become an attractive option in genome and transcriptome studies (*Au et al., 2013*; *Sharon et al., 2013*; *Xu et al., 2015*). In the present study, we performed SMRT sequencing to generate large-scale full-length or long-read transcripts from *I. batatas* and *I. trifida*, respectively. Comprehensive intraspecific and interspecific sequence analyses were conducted, which has provided a valuable resource for the research community to exploit the origin of *I. batatas*.

MATERIALS & METHODS

Plant material and RNA preparation

Xushu18, one of the most widely cultivated *I. batatas* varieties in China, was selected for transcriptome sequencing in this study. Eight tissues of young leaves, mature leaves, apical shoots, mature stems, fibrous roots, initiating tuberous roots, expanding tuberous roots, and mature tuberous roots from one individual were collected and pooled together in approximately equivalent weights (Figs. 1A–1H). Similarly, tissues of young leaves, mature leaves, shoots, stems, and roots of a diploid *I. trifida* plant were collected and pooled. Collected samples were frozen in liquid nitrogen immediately after collection and stored at −80 °C until use.

Total RNAs were extracted using Tiangen RNA preparation kits (Tiangen Biotech, Beijing, China) following the provided protocol. RNA quality and quantity were determined using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). Qualified RNA samples were subsequently used in constructing PacBio cDNA or RNA-seq libraries.

PacBio cDNA library construction and SMRT sequencing

cDNA was synthesized using a SMARTer PCR cDNA Synthesis Kit, optimized for preparing full-length cDNA (Takara Clontech Biotech, Dalian, China). Size fractionation and selection (1–2 kb, 2–3 kb, and >3 kb) were performed using the BluePippin^{TM} Size Selection System (Sage Science, Beverly, MA, USA). The SMRT bell libraries were constructed with the Pacific Biosciences DNA Template Prep Kit 2.0. SMRT sequencing was then performed on the Pacific Bioscience RS II platform using the provided protocol.

Illumina RNA-Seq library construction and sequencing

The RNA-Seq libraries were constructed using a NEBNext® Ultra^{TM} RNA Library Prep Kit for Illumina® (NEB, Beverly, MA, USA), following the manufacturer's protocol.

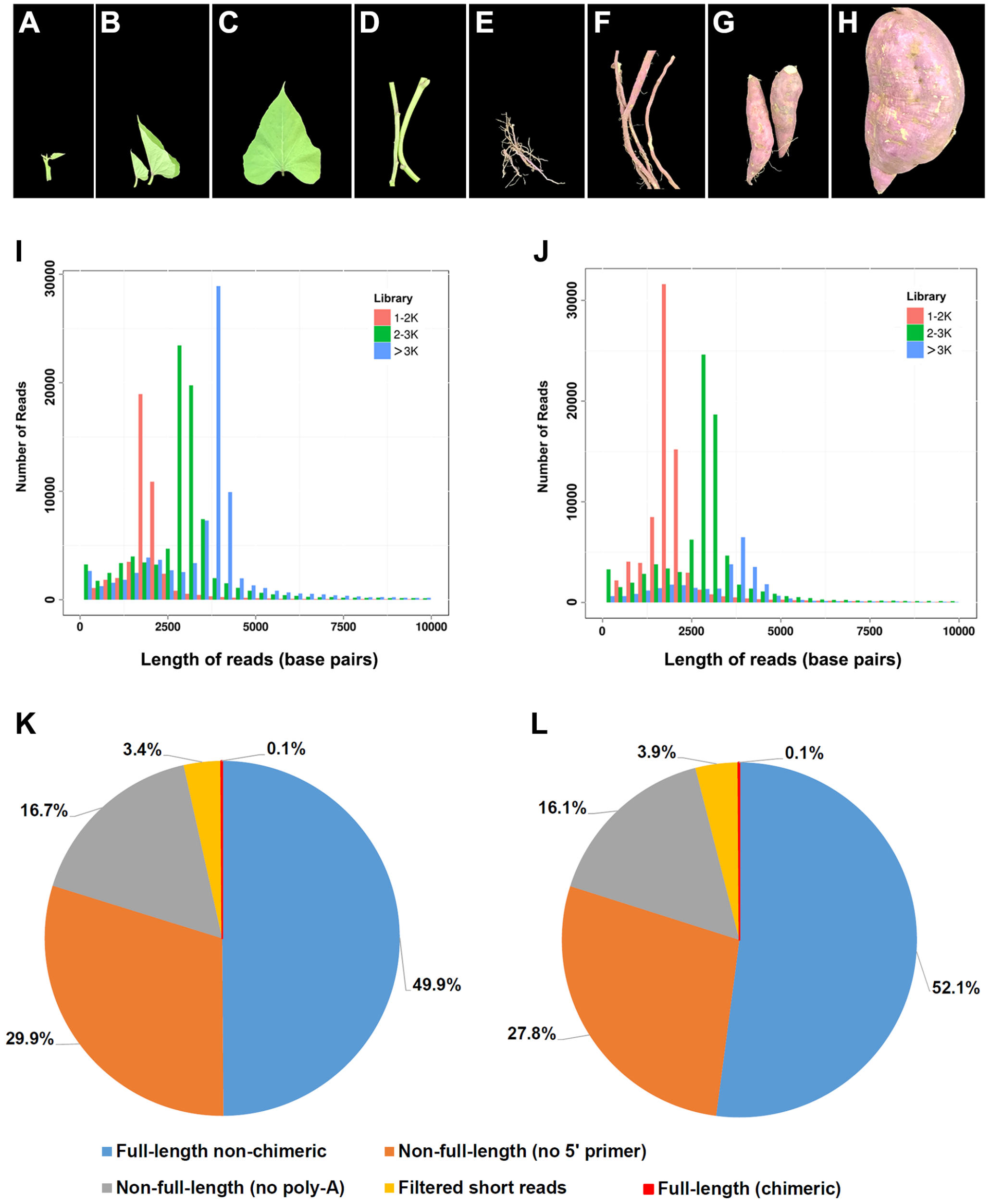

Figure 1 Plant materials used in this study and summary of PacBio RS II single-molecule real-time (SMRT) sequencing. (A–H) Photos showing the developmental stages and overall morphology of eight tissues in *I. batatas* used for SMRT sequencing in this study. (A) Young leaves; (B) mature leaves; (C) apical shoots; (D) mature stems; (E) fibrous roots; (F) initiating tuberous roots; (G) expanding tuberous roots; (H) mature tuberous roots. The photos were adopted from our previous report (*Ding et al., 2017*). Number and length distributions of 220,035 reads in *I. batatas* (I) and 195,188 reads in *I. trifida* (J) from different PacBio libraries (fractionated size: 1–2, 2–3, >3 kb); Proportion of different types of PacBio reads in *I. batatas* (K) and *I. trifida* (L).

Full-size DOI: 10.7717/peerj.7933/fig-1

Qualified libraries were applied to transcriptome sequencing using an Illumina Hiseq 2500 (Illumina, San Diego, CA, USA) to generate 150-bp paired-end sequence reads (2 × 150 bp). High-throughput sequencing reported in this study was performed in the Biomarker Technology Co. (Beijing, China).

Quality filtering and error correction of SMRT long reads

The SMRT subreads were filtered using the standard protocols in the SMRT Analysis software suite (http://www.pacificbiosciences.com), and reads of insert (ROIs) were obtained using the standard protocols in the SMRT Analysis software suite (parameters: minFullPass=0, minPredictedAccuracy=75). After examining for poly(A)signals and 5′ and 3′ adaptors, full-length and non-full-length cDNA reads were recognized. Consensus isoforms were identified using the algorithm of iterative clustering for error correction and further polished to obtain high-quality consensus isoforms. The raw Illumina reads were filtered to remove adaptor sequences, ambiguous reads with 'N' bases, and low-quality reads. Afterward, error correction of low-quality isoforms was conducted using the Illumina reads with the software proovread 2.13.841 (parameters: –coverage=50 –overwrite, –no-sampling) (*Hackl et al., 2014*). Redundant isoforms were then removed to generate a high-quality transcript dataset for each species (i.e., Ib53861 for *I. batatas* and It51184 for *I. trifida*, respectively) using the program CD-HIT 4.6.142 (parameters: -c 0.99 -T 6 -G 0 -aL 0.90 -AL 100 -aS 0.99 -AS 30 -o) (*Li & Godzik, 2006*).

Functional assignment of transcripts

Functional annotations were conducted by using BLASTX (cutoff E-value $\leq$ 1e−5) against different protein and nucleotide databases of COG (clusters of orthologous Groups; https://www.ncbi.nlm.nih.gov/COG/), GO (gene ontology; http://geneontology.org/), KEGG (kyoto encyclopedia of genes and genomes; https://www.kegg.jp/), Pfam (a database of conserved protein families or domains; http://pfam.xfam.org/), Swiss-prot (a manually annotated, non-redundant protein database; https://www.uniprot.org/), TrEMBL (an automatically annotated protein database; https://www.uniprot.org/), and NR (NCBI non-redundant proteins; https://www.ncbi.nlm.nih.gov/). For each transcript in each database searching, the functional information of the best matched sequence was assigned to the query transcript.

Predictions of open reading frames and simple sequence repeats

To predict putative open reading frames (ORFs) in transcripts, we used the package TransDecoder v2.0.1 (https://transdecoder.github.io/) to define coding sequences (CDS). The predicted CDS were searched and confirmed by BLASTX (E-value ≤1e−5) against the protein databases of NR, SWISS-PROT, and KEGG. Those transcripts containing complete ORFs as well as 5′- and 3′-UTR (untranslated regions) were designated as full-length transcripts. To identify putative simple sequence repeats (SSRs) in our sequences, the tool MISA (MIcroSAtellite identification tool; http://pgrc.ipk-gatersleben.de/misa) was employed. Only transcripts that were ≥500 bp in size were included in SSR detection.

Identification of transcription factor gene families

This was done according to our previous publication (*Ding et al., 2017*). Briefly, for each transcription factor gene family, the Hidden Markov Model (HMM) profile of the Pfam domain (when available) was downloaded from the Pfam database (http://pfam.xfam.org) and used as a query to survey all predicted proteins out of our transcript datasets using HMMER (http://www.hmmer.org). When no HMM profile was available for a gene family, all protein sequences belonging to the gene family in *A. thaliana* were downloaded (http://www.arabidopsis.org) and used as query sequences to search for our predicted protein datasets using BLASTP (*E*-value ≤1e−10). One redundant sequence was removed if two proteins shared the identity of amino acids equal to or larger than 97%. All identified non-redundant proteins were confirmed the existence of featured domains by searching the NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). The confirmed protein sequences as well as their corresponding transcripts were compiled (Only those gene families containing more than 10 members in at least one of our transcriptomes were presented).

Prediction of long non-coding RNAs

To sort non-coding RNAs from putative protein-coding ones, we employed each of four computational approaches including CPC (*Kong et al., 2007*), CNCI (*Liang et al., 2013*), Pfam (*Finn et al., 2016*), and CPAT (*Wang et al., 2013*). Putative protein-coding RNAs were filtered out using a minimum length and exon number threshold according the instructions of programs. For each species, the intersection of the four resulting lists were obtained as final lncRNA candidates.

Identification and validation of alternative splicing

To identify alternative splicing (AS) events, all transcripts of Ib53861 and It51184 were mapped to the the genomic contigs in *I. batatas* (*Yang et al., 2017*) and *I. trifida* (*Wu et al., 2018*), respectively, by using the program GMAP (*Wu & Watanabe, 2005*). The tool AStalavista v3.2 was employed to identify putative AS events (*Foissac & Sammeth, 2007*). Subsequently, 16 of AS events were selected and 10 of them were successfully confirmed by RT-PCR. Total RNA was isolated from the eight tissues in a *I. batatas* cultivar (*Xushu22*) as described above. The cDNA was synthesized using a cDNA Synthesis Kit (ProbeGene, China) and used as the template for PCR amplification. Afterward, PCR products were visualized in agarose gel.

RESULTS

SMRT sequencing and generation of full-length transcriptomes

To obtain large-scale long-read transcripts for *I. batatas* and *I. trifida*, respectively, SMRT sequencing was performed using a Pacific RSII sequencing platform. Eight different tissues collected from a single plant of each species were pooled and used in mRNA extraction. Three size-fractionated, full-length cDNA libraries were constructed and subsequently sequenced in four SMRT cells (Figs. 1I and 1J; 1–2 kb for one cell, 2–3 kb for two cells, and >3 kb for one cell). In *I. batatas*, we obtained 220,035 reads of the insert (total

Table 1 Summary of PacBio sequencing in this study.

	I. batatas	*I. trifida*
Reads of insert of PacBio sequencing	220,035	195,188
Bases of insert of PacBio sequencing (bp)	701,923,565	527,497,043
Reads of Illumina sequencing for correction	71,360,785	39,372,131
Bases of Illumina sequencing for correction (bp)	17,972,706,252	11,772,267,169
Number of non-full-length PacBio reads	102,510	85,680
Number of full-length non-chimeric PacBio reads	109,814	101,630
Average length of full-length non-chimeric PacBio reads (bp)	8,641	8,488
Number of non-redundant transcripts after correction	53,861	51,184
N50 of non-redundant transcripts after correction (bp)	2,933	2,642
Mean of non-redundant transcripts after correction (bp)	2,421	2,190
Number of non-redundant full-length transcripts after correction	34,963	33,637

bases: 701,923,565), which included 49.9% of full-length non-chimeric and 46.6% of non-full-length reads (Table 1, Fig. 1K), whereas in *I. trifida*, 195,188 reads of the insert (total bases: 527,497,043) were generated, of which 52.1% and 43.9% were full-length non-chimeric and non-full-length reads, respectively (Table 1, Fig. 1L).

Given that SMRT sequencing generates a high error rate, it is necessary to perform error correction, which includes self-correction by iterative clustering of circular-consensus reads and correction with high-quality Illumina short reads. To this end, cDNA libraries were prepared from the same samples that were used for SMRT sequencing, and deep RNA sequencing was conducted using an Illumina Hiseq2500 platform. A total of 71,360,785 and 39,372,131 clean reads (total bases: 17,972,706,252 and 11,772,267,169, respectively) were obtained and used to correct the SMRT reads in *I. batatas* and *I. trifida*, respectively (Table 1). After error correction, redundant transcripts were removed. Finally, we obtained 53,861 transcripts for *I. batatas* (named as Ib53861; N50: 2,933 bp; mean: 2,421 bp) and 51,184 for *I. trifida* (named as It51184; N50: 2,642 bp; mean: 2,190 bp). Those transcripts containing complete coding sequences (CDSs) as well as 5′- and 3′-UTR (untranslated regions) were defined as full-length transcripts. Approximately 34,963 and 33,637 full-length transcripts were identified for *I. batatas* (named as Ib34963) and *I. trifida* (named as It33637), respectively (File S1).

Basic sequence analysis of the full-length transcriptomes

The transcripts of Ib53861 and It51184 were functionally assigned and classified according to sequence similarities using BLASTx or tBLASTx (*E*-value $\leq$1e$-$5) against different protein and nucleotide databases. Overall, we successfully identified homologous sequences for 97.25% of Ib53861 and 97.34% of It51184 in the public databases, and the rates of successful validation in a single database ranged from 41.67% to 96.46% (File S2). These results indicate that most of the genes in our datasets are truly transcribed sequences in *I. batatas* and/or *I. trifida*. Furthermore, from the datasets of Ib53861 and It51184, 104,540/94,174 open reading frames (File S1), 25,315/27,090 simple sequence repeats

(Files S3–S5), 1,401/1,457 transcription factors (Files S6–S8), 1,656/1,389 long non-coding RNAs (Files S9 and S10), and 5,251/8,901 alternative splicing events (Files S11–S14) were identified. These data provide fundamental information for functional genomics study and molecular breeding in *I. batatas* and comparative biology study between *I. batatas* and *I. trifida*.

Analysis of long non-coding RNA

Recent studies have shown that lncRNAs act as key regulators in a wide range of biological processes. In the present study, we *in silico* identified 1,656 and 1,389 candidate lncRNAs out of Ib53861 and It51184, respectively (Figs. 2A and 2B; Files S9 and S10). Amongst, 421 *I. batatas* and 355 *I. trifida* transcripts could be recognized as sense, intergenic, intronic, or antisense lncRNAs (Fig. 2C). Notably, there were only 344 common candidate lncRNAs (i.e., homologs in sequences; cutoff: identity >200 bp & >90%) between the identified 1,656 *I. batatas* and 1,389 *I. trifida* transcripts, suggesting remarkable divergence in lncRNA biogenesis and thus their regulatory mechanisms between two species (Fig. 2D). These data suggest that different lncRNA members may be involved in different tissue/organ developmental processes in *I. batatas*.

Analysis of Alternative splicing

Alternative splicing (AS) is a posttranscriptional regulatory mechanism to increase transcriptome diversity, yet little is known about its roles in the development of tuberous root and the evolution of *I. batatas*. In the present study, we identified 5,251 and 8,901 AS events out of 10,562 and 17,826 transcript isoforms in *I. batatas* and *I. trifida*, respectively (Table 2; Files S11–S14). The AS events were divided into five major types: intron retention (IR), alternative 3′ splice site (A3SS), alternative 5′ splice site (A5SS), exon skipping (ES), and mutually exclusive exon (MEX; Fig. 3A). The proportion of each AS type was comparable between *I. batatas* and *I. trifida* and the majority of AS events were IR in either species (Fig. 3B). Overall, the alternatively spliced isoforms accounted for 32.34% or 38.54% of all isoforms successfully mapped to *I. batatas* scaffolds or *I. trifida* genome (Table 2), which should have largely increased the complexity of transcriptomes in either species. Notably, 37% of the alternatively spliced isoforms in *I. batatas* were not alternatively spliced or not detected in *I. trifida* and so were 63% of the alternatively spliced isoforms in *I. trifida*, suggesting substantial divergence in AS biogenesis and thus their regulatory mechanisms between two species (Fig. 3C). The isoform number per AS event ranged from 2 to 35 (mean, 4.98) in *I. batatas* and from 2 to 46 (mean, 4.55) in *I. trifida* (Table 2; Fig. 3D). In total, 2,074 loci in *I. batatas* and 3,640 in *I. trifida* were involved in the detected AS events (Table 2). The maximal number of AS events per locus was 45 (mean, 2.57) in *I. batatas* and 38 (mean, 2.45) in *I. trifida* (Table 2; Fig. 3E).

To assess our large-scale predictions of AS events, we manually examined 40 genes that were predicted as containing AS events and found 8 of them were likely false candidates. We then designed primers to examine 16 AS events, each of which located in one gene, by RT-PCR across eight tissues of an *I. batatas* variety (*Xushu22*), and successfully confirmed 10 of them (Figs. 3F and 3G). According these results, we concluded that at least 50% of

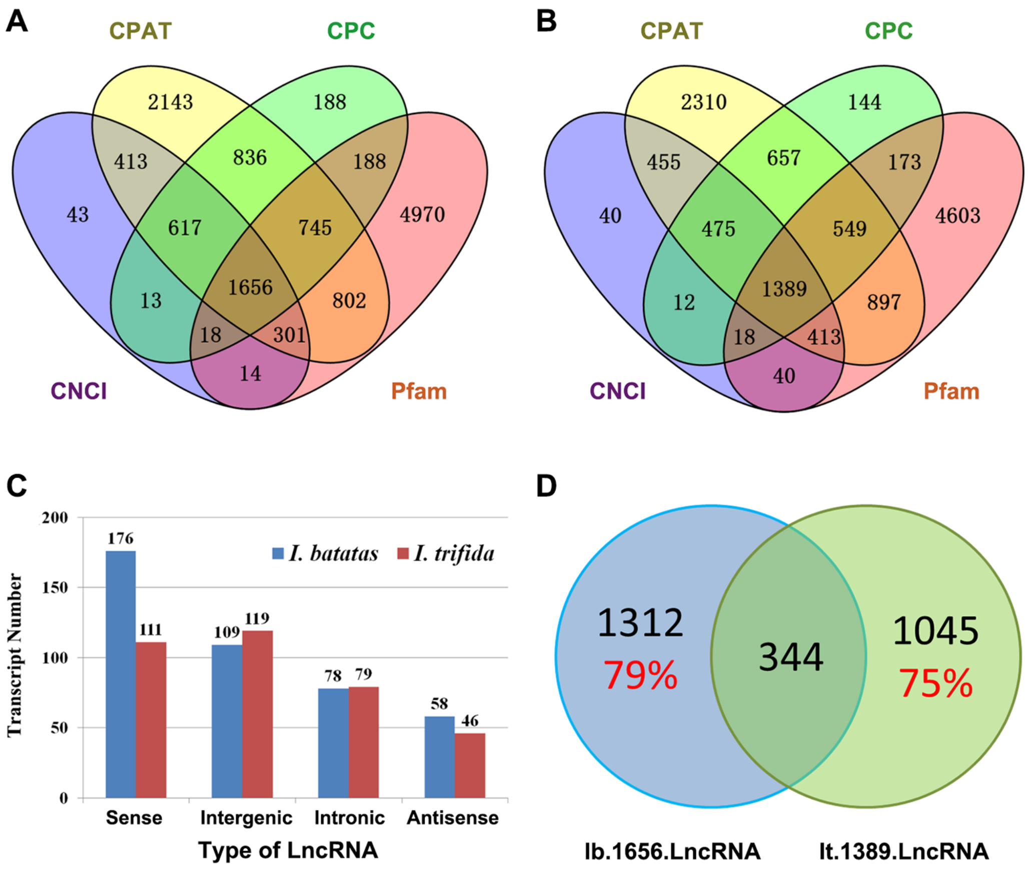

Figure 2 Analysis of putative long noncoding RNA (lncRNA). Venn Diagrams of lncRNAs predicted from Ib53861 (A) and It51184 (B) by four programs (CPAT, CPC, CNCI, and Pfam). (C) Types and numbers of lncRNAs that could be clarified in our analysis. (D) Homolog relationship of predicted lncRNAs between *I. batatas* and *I. trifida*.

Full-size DOI: 10.7717/peerj.7933/fig-2

our AS predictions were valid. Given that we only examined one of multiple AS events in each gene and only in one *I. batatas* variety, our data should be underestimated. Therefore, our large-scale AS analysis has provided a useful resource for studying biological functions of transcript isoforms and the regulatory mechanism of alternative splicing during the evolution of *I. batatas*.

For example, starch-branching enzymes (EC 2.4.1.18) are one of key enzymes involved in plant starch biosynthesis and sugar metabolism (*Zeeman, Kossmann & Smith, 2010*). In our analysis, we detected multiple AS events (i.e., one ES and one IR events) in a putative *I. batatas* starch-branching enzyme I and verified two AS isoforms, whose expression changed over different tissues (Fig. 3G, Gene01). In aboveground tissues (i.e., T01 to T04) and fibrous roots (i.e., T05), the two isoforms were expressed at a similar level; whereas in tuberous roots (i.e., T06 to T08), the smaller isoform were specifically expressed (Fig. 3G, Gene01). Plant alpha-glucan phosphorylases, also named as starch phosphorylase (EC 2.4.1.1), are another important family of enzymes involved in carbohydrate metabolism (*Rathore et al., 2009*). Our results revealed distinct splicing mechanisms existed between

Table 2 **Summary of alternative splicing analysis.**

	I. batatas	*I. trfida*
Number of isoforms of the datasets	53,861	51,184
Number of isoforms mapped to genome sequences	32,660	46,249
Number of isoforms (with multiple involvements) in AS events	26,146	40,473
Number of isoforms (with one involvement) in AS events	10,562	17,826
Number of detected AS events	5,251	8,901
Maximal number of isoforms in a single AS event	35	46
Mean number of isoforms per AS event	4.98	4.55
Number of loci occuring AS events	2,047	3,640
Maximal number of AS events in a single locus	45	38
Mean number of AS events per locus	2.57	2.45
Mean number of isoforms (with one involvement) per locus	5.16	4.90
Proportion of isofroms undergone AS	32.34%	38.54%
Number of estimated loci in the datasets	10,439	10,452

aboveground and belowground tissues in the examined *I. batatas* alpha-glucan phosphorylase (Fig. 3G, Gene02). In addition, divergent gene-expression and splicing patterns were also observed in other investigated genes including a neutral invertase, an E3 ubiquitin-protein ligase, a pentatricopeptide repeat-containing protein, and a few ABC transporters (Fig. 3G, Gene03–10). These data revealed that alternative splicing and thus transcriptome regulation might play important roles during the development of tuberous roots in *I. batatas*.

DISCUSSION

Understanding the genetic basis and evolutionary scenario underlying agronomically important traits is one of central research themes in the hexaploid crop *I. batatas*. However, achieving this goal is doomed to be challenging because of the complexity of its genome structure (*Isobe, Shirasawa & Hirakawa, 2017*). In the present study, we applied a hybrid sequencing approach to generate and analyze large-scale full-length or long-read transcripts and their expression profiles in *I. batatas*. Our study would be beneficial to the *I. batatas* research community at least in the following aspects: gene cloning, gene family analysis, development of cDNA-derived marker for breeding, gene model prediction, genome assembly, and study of genetic variation within or among species. For example, we have demonstrated an example of fast gene cloning and gene family analysis basing on our transcriptome datasets (*Ding et al., 2017*). Overall, our study has provided a fundamental resource for functional genomics study in *I. batatas*, which would certainly facilitate genetic dissections of the origin of tuberous root as well as other traits.

AS commonly occurs in eukaryotes. In humans, more than 90% of genes were found to be alternatively spliced and the predominant AS type was exon-skipping (*Wang et al., 2008*). In higher plants, the AS frequency in intron-containing genes

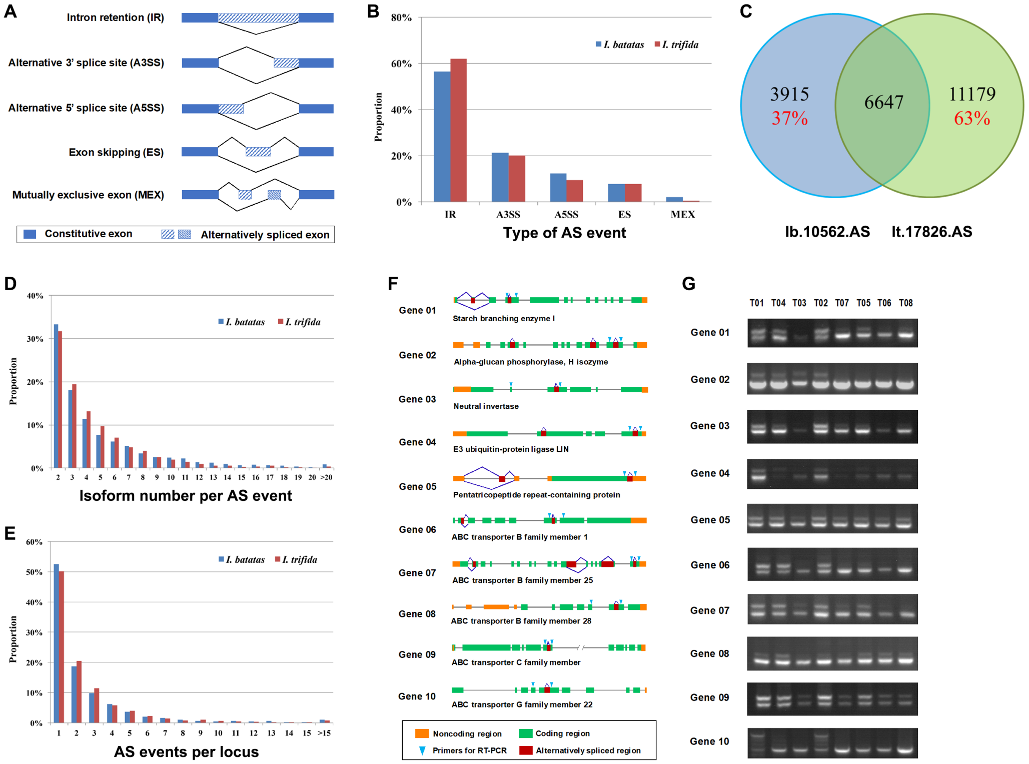

Figure 3 Analysis and validation of alternative splicing (AS). (A) Diagrams showing five major AS types. (B) Proportions of major AS types predicted out of the dataset Ib53861 and It51184. (C) Homolog relationship of isoforms carrying putative AS events between *I. batatas* and *I. trifida*. (D) Proportion distribution of isoform number per AS event in *I. batatas* and *I. trifida*. (E) Proportion distribution of AS events per locus in *I. batatas* and *I. trifida*. (F) Diagram and (G) RT-PCR validation of AS events in ten *I. batatas* genes.

Full-size DOI: 10.7717/peerj.7933/fig-3

approximately ranged from 33% to 60% with intron retention as the major type (*Filichkin et al., 2010*; *Zhang et al., 2010*; *Shen et al., 2014*; *Thatcher & Li, 2014*). In our study, we observed an overall AS frequency of 32.34% in *I. batatas* isoforms (Table 2). Considering about 30.03% of isoforms contained a single exon in our dataset, the AS frequency in intron-containing isoforms in *I. batatas* was approximately 46.22%. The estimated AS frequency in intron-containing isoforms in *I. trifida* was 51.18%, a little bit higher than that of *I. batatas*. The major AS type was intron retention in either *I. batatas* or *I. trifida*, similar as observed in other plants. These data highlighted the prevalence of AS in both *I. batatas* and *I. trifida*, which would certainly increase the complexity of their transcriptomes. In addition, we also examined the AS pattern across eight tissues in 10 *I. batatas* genes and found that many isoforms exhibited a tissue-specific expression pattern (Fig. 3G). These results imply that the generation of AS isoforms in a tissue-dependent manner have contributed substantially to organ/tissue development and species evolution in *I. batatas*.

AS and gene/genome duplication are two fundamental biological processes contributing to transcriptome and proteome diversity. The relationship between these two evolutionary mechanisms remains debatable. Some studies have reported that the AS frequency decreased after gene duplication and genome polyploidization (*Kopelman, Lancet & Yanai, 2005*; *Su et al., 2006*). In contrast, some other reports argued that the evolutionary relationship between AS and gene/genome duplication was more complex and must be cautiously anticipated (*Lin et al., 2008*; *Roux & Robinsonrechavi, 2011*; *Iñiguez & Hernández, 2017*). In this study, our transcriptome-wide AS analysis revealed comparable AS patterns between *I. batatas* and *I. trifida*, in terms of mean number of isoforms per AS event or per locus, mean number of AS events per locus, and proportion of isoforms undergone AS (Table 2; Fig. 3). These data showed that the overall AS frequency (not between specific duplicated gene pairs) was not evidently decreased after the genome hexaploidization in *I. batatas*.

CONCLUSIONS

Although *I. batatas* is a global crop of great agronomic importance, advances in its functional genomics study and molecular breeding remain limited because of the complexity of its genome. Here we report the first collections and analyses of large-scale full-length or long-read transcripts in *I. batatas* and its putative diploid ancestor *I. trifida* using single-molecule real-time sequencing. By performing comprehensive intraspecific and interspecific sequence analyses, we provide a valuable resource for genetic marker development, gene discovery, and gene function study in *I. batatas*, as well as comparative biology study between *I. batatas* and *I. trifida*. Furthermore, we analyzed transcriptome-wide long non-coding RNA and alternative splicing, which revealed tissue-specific-expressed transcript isoforms and the importance of transcriptome regulation during the speciation and domestication of *I. batatas*.

ACKNOWLEDGEMENTS

We would like to thank Prof. Dr. Daifu Ma and Prof. Dr. Dabing Zhang for their kind help in the project performance, and Ms. Xuan Shi, Ms. Ruyuan Wang, and Mr. Yunxiang Wu for their technical assistance.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was jointly supported by the National Natural Science Foundation of China (Grant No. 31771855), the National Sweet Potato Industry and Research System (Grant No. CARS-11-B-02), the startup funding from Fujian Agriculture and Forestry University, and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 31771855.
National Sweet Potato Industry and Research System: CARS-11-B-02.

Fujian Agriculture and Forestry University.
Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Na Ding performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Huihui Cui performed the experiments, analyzed the data, prepared figures and/or tables, approved the final draft.
- Ying Miao, Jun Tang, Qinghe Cao and Yonghai Luo conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The assembled sequences of full-length transcriptomes are available at DDBJ/EMBL/-GenBank:GHYO00000000.

Data Availability

The following information was supplied regarding data availability:

The PacBio SMRT reads and the Illumina short reads are available at the Genome Sequence Archive of Beijing Institute of Genomics, Chinese Academy of Sciences (https://bigd.big.ac.cn/gsa/browse/CRA000288).

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.7933#supplemental-information.

REFERENCES

Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, Bakel HV, Schadt EE, Reijopera RA, Underwood JG. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110(50)**:4821–4830 DOI 10.1073/pnas.1320101110.

Cao Q, Li A, Chen J, Sun Y, Tang J, Zhang A, Zhou Z, Zhao D, Ma D, Gao S. 2016. Transcriptome sequencing of the sweet potato progenitor (*Ipomoea Trifida* (H.B.K.) G. Don.) and discovery of drought tolerance genes. *Tropical Plant Biology* **9(2)**:63–72 DOI 10.1007/s12042-016-9162-7.

Ding N, Wang A, Zhang X, Wu Y, Wang R, Cui H, Huang R, Luo Y. 2017. Identification and analysis of glutathione S-transferase gene family in sweet potato reveal divergent

GST-mediated networks in aboveground and underground tissues in response to abiotic stresses. *BMC Plant Biology* **17(1)**:225 DOI 10.1186/s12870-017-1179-z.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D. 2010. Real-time DNA sequencing from single polymerase molecules. *Methods in Enzymology* **472(5910)**:431–455 DOI 10.1016/S0076-6879(10)72001-2.

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC. 2010. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research* **20(1)**:45–58 DOI 10.1101/gr.093302.109.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangradorvegas A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44(D1)**:D279–D285 DOI 10.1093/nar/gkv1344.

Foissac S, Sammeth M. 2007. Astalavista: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Research* **35**:297–299 DOI 10.1093/nar/gkm311.

Hackl T, Hedrich R, Schultz J, Förster F. 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30(21)**:3004–3011 DOI 10.1093/bioinformatics/btu392.

Iñiguez LP, Hernández G. 2017. The evolutionary relationship between alternative splicing and gene duplication. *Frontiers in Genetics* **8**:14 DOI 10.3389/fgene.2017.00014.

Isobe S, Shirasawa K, Hirakawa H. 2017. Challenges to genome sequence dissection in sweetpotato. *Breeding Science* **67(1)**:35–40 DOI 10.1270/jsbbs.16186.

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* **35**:345–349 DOI 10.1093/nar/gkm391.

Kopelman NM, Lancet D, Yanai I. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nature Genetics* **37(6)**:588–589 DOI 10.1038/ng1575.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, Mccombie WR, Jarvis ED. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology* **30(7)**:693–700 DOI 10.1038/nbt.2280.

Li W, Godzik A. 2006. Cd-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22(13)**:1658–1659 DOI 10.1093/bioinformatics/btl158.

Liang S, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Yi Z. 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research* **41(17)**:e166 DOI 10.1093/nar/gkt646.

Lin H, Ouyang S, Egan A, Nobuta K, Haas BJ, Zhu W, Gu X, Silva JC, Meyers BC, Buell CR. 2008. Characterization of paralogous protein families in rice. *BMC Plant Biology* **8(1)**:18 DOI 10.1186/1471-2229-8-18.

Magoon ML, Krishnan R, Vijaya BK. 1970. Cytological evidence on the origin of sweet potato. *Theoretical and Applied Genetics* **40(8)**:360–366 DOI 10.1007/BF00285415.

Martin FW. 1965. Incompatibility in the sweet potato. A review. *Economic Botany* **19(4)**:406–415 DOI 10.1007/BF02904812.

Nurit F, Don LB, Arthur V, Yanir K, Julio S, Evgenia L, Schnitzer PT, Adi DF, Amots H, Leviah A. 2013. Transcriptional profiling of sweetpotato (*Ipomoea batatas*) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. *BMC Genomics* **14(1)**:460 DOI 10.1186/1471-2164-14-460.

Ozias-Akins P, Jarret RL. 1994. Nuclear DNA content and ploidy levels in the genus ipomoea. *Journal of the American Society for Horticultural Science* **119(1)**:110–115 DOI 10.21273/JASHS.119.1.110.

Ponniah SK, Thimmapuram J, Bhide K, Kalavacharla VK, Manoharan M. 2017. Comparative analysis of the root transcriptomes of cultivated sweetpotato (*Ipomoea batatas* [L.] Lam) and its wild ancestor (*Ipomoea trifida* [Kunth] G. Don). *BMC Plant Biology* **17(1)**:9 DOI 10.1186/s12870-016-0950-x.

Rathore RS, Garg N, Garg S, Kumar A. 2009. Starch phosphorylase: role in starch metabolism and biotechnological applications. *Critical Reviews in Biotechnology* **29(3)**:214–224 DOI 10.1080/07388550902926063.

Roberts RJ, Carneiro MO, Schatz MC. 2013. The advantages of SMRT sequencing. *Genome Biology* **14**:405 DOI 10.1186/gb-2013-14-6-405.

Roux J, Robinsonrechavi M. 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Research* **21(3)**:357–363 DOI 10.1101/gr.113803.110.

Schafleitner R, Tincopa LR, Palomino O, Rossel G, Robles RF, Alagon R, Rivera C, Quispe C, Rojas L, Pacheco JA. 2010. A sweetpotato gene index established by de novo assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers. *BMC Genomics* **11(1)**:604 DOI 10.1186/1471-2164-11-604.

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K. 2002. Functional annotation of a full-length arabidopsis cDNA collection. *Science* **296(5565)**:141–145 DOI 10.1126/science.1071006.

Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology* **31(11)**:1009–1014 DOI 10.1038/nbt.2705.

Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, Ma Y, Liu T, Kong LA, Peng DL. 2014. Global dissection of alternative splicing in paleopolyploid soybean. *The Plant Cell* **26(3)**:996–1008 DOI 10.1105/tpc.114.122739.

Shoshi K, Kouji S, Toshifumi N, Nobuyuki K, Koji D, Naoki K, Junshi Y, Masahiro I, Hitomi Y, Hisako O. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* **301(5631)**:376–379 DOI 10.1126/science.1081288.

Srisuwan S, Sihachakr D, Siljak-Yakovlev S. 2006. The origin and evolution of sweet potato (*Ipomoea batatas* Lam.) and its wild relatives through the cytogenetic approaches. *Plant Science* **171**(**3**):424–433 DOI 10.1016/j.plantsci.2006.05.007.

Su Z, Wang J, Yu J, Huang X, Gu X. 2006. Evolution of alternative splicing after gene duplication. *Genome Research* **16**(**2**):182–189.

Thatcher SR, Li B. 2014. Genome-wide analysis of alternative splicing in zea mays: landscape and genetic regulation. *The Plant Cell* **26**(**9**):3472–3487 DOI 10.1105/tpc.114.130773.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**(**7221**):470–476 DOI 10.1038/nature07509.

Wang L, Park HJ, Dasari S, Wang S, Kocher J, Li W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41**(**6**):e74 DOI 10.1093/nar/gkt006.

Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y. 2010. De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics* **11**(**1**):726 DOI 10.1186/1471-2164-11-726.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**(**1**):57–63 DOI 10.1038/nrg2484.

Wu S, Lau KH, Cao Q, Hamilton JP, Sun H, Zhou C, Eserman L, Gemenet DC, Olukolu BA, Wang H, Crisovan E, Godden GT, Jiao C, Wang X, Kitavi M, Manrique-Carpintero N, Vaillancourt B, Wiegert-Rininger K, Yang X, Bao K, Schaff J, Kreuze J, Gruneberg W, Khan A, Ghislain M, Ma D, Jiang J, Mwanga ROM, Leebens-Mack J, Coin LJM, Yencho GC, Buell CR, Fei Z. 2018. Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nature Communications* **9**:4580 DOI 10.1038/s41467-018-06983-8.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**(**9**):1859–1875 DOI 10.1093/bioinformatics/bti310.

Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, Zhu Y, Ji A, Zhang B, Hu S. 2015. 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. *The Plant Journal* **82**(**6**):951–961 DOI 10.1111/tpj.12865.

Yang J, Moeinzadeh M, Kuhl H, Helmuth J, Peng X, Haas S, Liu G, Zheng J, Zhe S, Fan W. 2017. Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature Plants* **3**(**9**):696–703 DOI 10.1038/s41477-017-0002-z.

Zeeman SC, Kossmann J, Smith AM. 2010. Starch: its metabolism, evolution, and biotechnological modification in plants. *Annual Review of Plant Biology* **61**(**1**):209–234 DOI 10.1146/annurev-arplant-042809-112301.

Zhang G, Guo GX, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L. 2010. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Research* **20**(**5**):646–654 DOI 10.1101/gr.100677.109.