Genome-wide mapping of alternative splicing in the elite wheat cultivar Xiaoyan 6

Hude Mao (✉ mhd163com@163.com)

Northwest Agriculture and Forestry University   https://orcid.org/0000-0002-2484-9952

Jin Zhang

Northwest Agriculture and Forestry University

Lijian Guo

Northwest Agriculture and Forestry University

Qing Chi

Northwest Agriculture and Forestry University

Fangming Mei

Northwest Agriculture and Forestry University

Xiangli Liu

Northwest Agriculture and Forestry University

Huixian Zhao

Northwest Agriculture and Forestry University

Research article

1 **Genome-wide mapping of alternative splicing in the elite wheat**

2 **cultivar Xiaoyan 6**

3

4 Hude Mao[1a*], Jin Zhang[2a], Lijian Guo[2], Qing Chi[2], Fangming Mei[1], Xiangli Liu[2],

5 Huixian Zhao[2*]

6

7 [1] State Key Laboratory of Crop Stress Biology for Arid Areas, College of Plant

8 Protection, Northwest A & F University, Yangling, Shaanxi 712100, China

9 [2] State Key Laboratory of Crop Stress Biology for Arid Areas, College of Life

10 Sciences, Northwest A & F University, Yangling, Shaanxi 712100, China

11

12 [a] These authors contributed equally: Hude Mao and Jin Zhang

13

14 ***** Corresponding author.

15 Huixian Zhao

16 E-mail address: hxzhao212@nwafu.edu.cn

17 Hude Mao

18 E-mail address: mhd163com@163.com

19

20

21

22

1 **ABSTRACT**

2 **Background:** *Triticum aestivum*, or common wheat, is a globally important cereal

3 crop. Alternative splicing (AS) is an important post-transcriptional regulation

4 mechanism in higher eukaryotes, but the knowledge of AS in wheat is limited at

5 present.

6 **Results:** In this study, we performed AS mapping from the elite, high-yielding wheat

7 cultivar Xiaoyan 6. We identified 18,960 and 25,133 novel protein-coding genes and

8 transcript models that were not annotated in the draft genome of wheat cultivar

9 Chinese Spring. We also found that 18,868 genes were alternatively spliced, and

10 classified four major AS types, including alternative acceptor sites (AA), intron

11 retention (IR), alternative donor sites (AD), and exon skipping (ES). Further analysis

12 determined that differences in the AS frequency and AS types displayed a positive

13 correlation with exon number and intron length and were negatively correlated with

14 the GC content, but were not correlated with gene transcriptional level. Furthermore,

15 we identified 2,737 seed-specific AS genes in wheat cultivar Xiaoyan 6, and

16 demonstrated that the number of AS genes in the grain decreased with seed

17 development, suggesting that the earlier developmental stages are more important and

18 complex than are the later stages. In addition, the grain-specific AS genes that were

19 ubiquitously significantly up-regulated during seed development were associated with

20 metabolism-related pathways, indicating that AS of metabolism-related genes is

21 crucial for grain filling in wheat.

22 **Conclusions:** Through transcriptome analysis, over ten thousand AS genes were

23 identified and characterized in wheat cultivar Xiaoyan 6. Our study provides a

24 comprehensive view of AS in the wheat cultivar Xiaoyan 6 and provides new insights

25 into the complexity of AS in wheat. The data generated in this study provides a

26 foundation for further studies of seed development in wheat.

27 **Keywords:** wheat, Xiaoyan 6, RNA-seq, alternative splicing, seed development

INTRODUCTION

Alternative splicing (AS) is a widespread mechanism in both animal and plant species that significantly heightens the complexity of proteomes and transcriptomes, and regulates development and environmental response in higher eukaryotes [1]. The splicing process is performed by the spliceosome, which consists of five small nuclear RNAs (snRNAs) as well as 180 proteins with varying purposes [2]. The AS profile and transcriptome are both determined by the number of splicing factors in various cell types, environmental conditions, development stages, and tissues. Additionally, these alternative transcripts can regulate protein isoforms with different levels of amino acids and domain structure, affecting their capacity for interaction, localization, activity, stability, and ultimately, the proteome [3]. However, not all AS events are functional. It has been estimated that premature termination codons (PTC) occur in about one-third of AS events [4, 5], which are then possible targets for the nonsense-mediated mRNA decay system [6, 7].

AS takes part in a wide range of plant functions, such as flowering, stress response, circadian rhythm, and growth and development [8-10]. Recent analyses of *Arabidopsis thaliana*, cotton (*Gossypium raiimondi*), soybean (*Glycine max*), rice (*Oryza sativa*), and maize (*Zea mays*) have all demonstrated the presence of AS in these plants and estimate that alternative splicing happens in approximately 60% of intron-containing genes in *Arabidopsis*, approximately 40% in cotton, approximately 33% in rice, approximately 52% in soybean, and approximately 40% in maize [11-15]. AS events are primarily grouped as intron retention (IR), alternative acceptor sites (AA), exon skipping (ES), and alternative donor sites (AD), relative to the main transcript isoform [16]. Approximately 95% of genes are alternatively spliced in humans, of which, ES events are the most commonly occurring AS type [17]. In contrast, plant AS events predominantly involve IR [18, 5, 11, 13]. However, the ratios of the different AS types can vary between monocots and dicots [14].

1   AS is catalyzed by the spliceosome, and the selection of intron removal is strongly

2   regulated by consensus splice sequences, where most plant introns have a 3' AG and a

3   5' GU. Compared to exons that are GC-rich, introns are more inclined toward AU;

4   some difference occurs between dicots and monocots [16, 9]. In the past two decades,

5   several research projects have sought to determine the regulators affecting AS [19-21].

6   Indeed, many regulators have been identified, like changes to chromatin [22, 23], the

7   promoter responsible for transcription [24, 25], alterations in the environment of

8   chromatin [26], and the addition of various transcription characteristics or

9   coactivators to promoters [27]. However, the process of AS is complex, and the

10   factors affecting the inhibition and activation of various forms of mRNA splices

11   require further research.

12   Following the detection of the first AS in the endogenous gene of a mammal, coding

13   for calcitonin [28], there has been an increase in the identification of alternatively

14   spliced genes in humans (~35%) due to the alignment of expressed sequence tag (EST)

15   contigs to genomic DNA [29]. The incidence of AS in plants is approximated using a

16   low EST count [30]. The AS levels that have been detected have increased with

17   increases in coverage of EST/cDNA [31-35, 5]. High-throughput sequencing results

18   in a large dataset for significant analyses, and, along with improved technologies,

19   increases the ability to study AS. While many studies examining transcriptomes using

20   high-throughput sequencing have been conducted in plants, few have research AS

21   [11-15].

22   *Triticum aestivum* L., or common wheat, is a crop grown worldwide and accounts for

23   20 percent of the calories that humans eat [36]. To investigate the global

24   characteristics of AS in wheat, we executed paired-end RNA sequencing of 20

25   samples from the elite wheat cultivar Xiaoyan 6 using the Illumina HiSeq platform,

26   and generated ~ 1,445 million 125 bp pair-end reads. Wheat cultivar Xiaoyan 6 has

27   good quality grain, a high and stable yield, as well as drought tolerance, and was also

1 used as one of the most important foundation parents in wheat breeding programs in

2 China for about 20 years. Through read mapping and transcript assemblies, we

3 classified new AS events and genes in Xiaoyan 6. In this study, we first describe the

4 characteristics of the different AS type events in wheat. We then identify the genomic

5 features that influence AS, such as the number of exons, the length of the introns, the

6 level of transcription, and the content of GC. We also found that differences in AS are

7 correlated with differences with gene expression regarding splicing factors. Finally,

8 we discuss the distribution, expression patterns, and functional enrichments of seed-

9 specific AS genes.

10 **RESULTS**

11 **Mapping the transcriptome of wheat cultivar Xiaoyan 6**

12 To identify alternative splicing events in wheat cultivar Xiaoyan 6, we executed a

13 high-throughput RNA-seq on 20 samples using the Illumina HiSeq 2500 DNA

14 sequencing instrument. The samples included RNA extracted from root (R), stem (S),

15 and leaf (L) of wheat seedlings at the five-leaf stage, young spikes at the early booting

16 stage (YS5), spikes at the heading stage (YS15), flag leaf at the heading stage (FL),

17 and grains at 5, 10, 15, and 20 days post-anthesis (GR5, GR10, GR15, and GR20,

18 respectively) (Additional file 1: Table S1). The Illumina sequencing generated nearly

19 1,445 million paired-end reads (~ 145Gb; 125-nt read length). We trimmed the

20 adaptor sequences and filtered out the low-quality reads, there remained ~ 1,387

21 million high-quality reads and an average per-sample read of 69.4 million reads (~

22 6.94 Gb) (Fig. 1A; Additional file 1: Table S1). We then mapped the reads to the

23 genome sequence of *T. aestivum cv*. Chinese Spring, which was used as a reference

24 (IWGSC RefSeq v1.0), after which we assembled the transcripts and performed

25 analyses of the gene expression and differential isoform via the Tuxedo RNA-seq data

26 analysis pipeline [37, 38]. The analysis yielded ~ 60 to 75% overall alignment rates

1 for the 20 samples (Additional file 1: Table S1). There were approximately 95%

2 unique reads, which were only mapped once to a particular genome, which indicates

3 high-quality sequencing (Fig. 1A). The number of aligned reads for a certain

4 chromosome displayed a high correlation with the total length of the gene in this

5 chromosome (Fig. 1B), implying extensive chromosome coverage. Additional file 2:

6 Figure S1 displays the distribution of the reads along each chromosome in the

7 windows of 1 kb. Additional examination demonstrated that most mapped reads were

8 found in exonic regions, while a smaller number of the reads were mapped to non-

9 gene or intronic regions (Fig. 1C). In addition, approximately 78% of the genes have

10 at least 60% coverage by mapped reads (Fig. 1D). The aligned reads from intronic and

11 exonic areas were then used to understand gene expression, and to determine the AS

12 events for the genes.

13 Along with the assemblies of our transcripts, the combined Xiaoyan 6 gene annotation

14 has ~ 120,770 transcripts among 103,480 gene loci. From these, ~ 18,960 (9%) and

15 25,133 (24%) were transcript and gene models that had been previously unannotated

16 (Fig. 1E). Each of these novel genes was used as a query in BLASTX searches (E-

17 value cutoff: 1e-10) and compared to databases of reference sequences that are

18 available to the public. This search found these transcripts as encoding putative

19 proteins having various characteristics, such as receptor-like kinase (XLOC_000629),

20 MYB-like protein (XLOC_002308), F-box protein (XLOC_000122), NBS-LRR

21 superfamily (XLOC_021305), NAC domain-containing protein (XLOC_023217), and

22 ABC transporter family (XLOC_081889) (Additional file 3: Table S2).

23 We then performed a global inspection of the isoform-level expression using the

24 CummeRbund program [37]. The isoform-level expression's density level ($\log_{10}$

25 FPKM) displayed a normal distribution across the 20 samples (Fig. 2A), with little

26 systematic bias. In addition, box plots, heatmaps, and pairwise scatter plots of the

27 isoform-level expression showed the higher correlations between each sample

1 replicate pair (Fig. 2B, C; Additional file 4: Figure S2). To better understand the

2 relationships between different transcriptomes, we performed principal component

3 analysis (PCA) based on the expression values of all identified genes in the 20

4 samples, which shows the similarities in development as well as the transcriptional

5 signatures graphically. The first part (49.6% of the variance explained) grouped the

6 samples based on their tissue, as well as separating developing seeds from young

7 spikes, with seedling roots, seedling stems, seedling leaves, and flag leaves located in

8 between (Fig. 2D). The second component (24.5% of the variance explained)

9 discriminated between different developmental stages of these tissues (Fig. 2D).

10 Together, these analyses demonstrated the robustness of the wheat cultivar Xiaoyan 6

11 transcriptome data at isoform-level resolution.

12 **High-level detection of AS genes in the wheat genome**

13 The AS events were identified using Alternative Splicing Transcriptional Landscape

14 Visualization (ASTALAVISTA) tool [39]. Of the expressed genes, 18,868 genes

15 (18.23%) were predicted to be subject to AS. Our set of transcripts was validated

16 using an RT-PCR alternative splicing panel, where we examined 13 genes, all

17 containing AS events, and which were either obtained from the *T. aestivum* reference

18 genome, or newly found in this study (Additional file 5: Figure S3). These findings

19 demonstrate the ability of RNA-seq to identify new AS events. Within each tissue, the

20 number of AS genes greatly varied. More AS genes were found in the stems (S) and

21 leaves (L) of wheat seedlings at the five-leaf stage, flag leaf (FL), and spike at the

22 heading stage (YS15), while the number of AS genes were less in other tissues/organs

23 (Additional file 6: Figure S4A). This inconsistency could be due to the total number

24 of genes expressed in each singular tissue. We performed a correlation analysis of the

25 samples, which demonstrated that the gene number of the AS displayed a high

26 correlation with the gene number that was expressed (Additional file 6: Figure S4B),

27 after which we determined the AS gene ratio (number of AS genes/total number of

1   expressed genes) to remove any bias in the number of expressed genes. We also

2   determined that AS gene ratios were higher in young stems, young leaves, flag leaves,

3   and spike at the heading stage (Additional file 6: Figure S4C). Further analysis

4   suggested that the AS frequency might be associated with gene expression levels

5   (Additional file 6: Figure S4D). We also identified individual protein isoform domains

6   in each gene (Fig. 3A). Our findings showed that compared to non-AS genes, AS

7   genes contained more domains, which were consistent with previous research

8   performed on both *Arabidopsis* and rice [40]. After comparing the structural features

9   of non-AS genes and AS genes, we determined that AS genes were longer than the

10   non-AS genes (Fig. 3B), the exons of AS genes were also longer than those of non-AS

11   genes (Fig. 3C), and AS genes possess more exons (Fig. 3D), indicating that gene

12   structure might be the cause of the difference in domain number observed in non-AS

13   genes and AS genes.

14   Among the 18,868 AS genes, 709 were alternatively spliced in each sample examined

15   by our research (Fig. 3E). To identify which biological processes the alternatively

16   spliced genes could play a role in, we studied Gene Ontology (GO) term enrichment

17   using all *T. aestivum* genes as controls. Our findings demonstrated that AS genes were

18   highly expressed in mechanisms involving metabolism and development, and RNA

19   splicing-related processes were also enriched, such as RNA processing (GO:0006396,

20   P = 1.05E-04), RNA splicing (GO:0008380, P = 1.12E-04), and RNA splicing

21   regulation (GO:0043484, P = 1.62E-04) (Fig. 3F).

22   **Frequency and features of different types of AS events**

23   AS types were also categorized with the ASTALAVISTA [39]. All 90,546 AS events

24   were identified and classified into: (1) IR (intron retention), (2) AA (alternative

25   acceptor sites), (3) AD (alternative donor sites), (4) ES (exon skipping), as well as (5)

26   alternate incidents. The "alternate incidents" are comprised of AS events involving

1   duplicated IR, AD, AA, and ES events, or a mix of some. Among the different AS

2   types, IR represented 32.90% of the sum and was most commonly found. The next

3   most common was AA (31.96%), AD (17.76%), and ES (9.51%) events (Fig. 4A).

4   Approximately 8% of the events were classified as complex AS events. These results

5   agreed with previous observations in *Arabidopsis* [11, 13]. The four AS types varied

6   remarkably in the individual tissues (Fig. 4B). Interestingly, we found that IR was the

7   predominant AS type in the FL, GR5, GR10, GR15, and GR20 samples, while in the

8   R, S, LYS5, and YS15 samples, AA was the predominant AS type (Fig. 4B;

9   Additional file 7: Figure S5).

10   A recent study of AS in soybean indicated that ratios of various AS types could differ

11   between dicots and monocots [14]. We analyzed the AS samples, along with the

12   relative proportion of AS types in four reference monocots to test the following:

13   wheat (this study), *B. distachyon* (Bd21v1.2), rice (*O. sativa*; MSU6.0), and maize (*Z.*

14   *mays*; AGPv3). Using the ASTALAVISTA program, we also analyzed four dicots as

15   controls: *Arabidopsis* (TAIR10), potato (*S. tuberosum*; SolTub3.0), soybean (*G. max*

16   Assembly, v1.0), and poplar (*P. trichocarpa*; JGI2.0). The results showed that the

17   proportions of various AS types and the landscapes of wheat were similar to maize,

18   rice, and *B. distachyon* (Fig. 4C). We found few differences in the proportion of

19   various AS types between dicots and monocots (Fig. 4C). Among the monocots and

20   dicots that we examined, IR events were the most prevalent AS type; AA types and

21   AD types were the next two. The least common AS type among plants were ES events

22   (Fig. 4C). These findings demonstrate that the mechanisms regulating intron

23   processing and AS are conserved across dicots and monocots, lending support to the

24   idea that IR events are more common in plants than in animals [9].

25   The use of specific locations to splice was an integral component of AS. Of the AS

26   events we found in our research, 78,901 were GT-AG splice locations (87.14%),

27   3,886 were GC-AG splice locations (4.30%), 235 were AT-AC splice locations

1   (0.26%), and 7,524 were other splice locations (8.31%). The GT-AG location was the

2   most common of the four kinds of AS, while GC-AG was the second most common

3   (Additional file 8: Figure S6), results that are consistent with those of other species

4   [41, 42, 11]. Although the GT-AG splice site was also the most common in the

5   different AS types, proportions of use for the splice sites differed among the four

6   kinds of AS. In ES, the proportion of GT-AG was ~ 97%, which was higher than in IR

7   (80.18%), AA (88.38%), and AD (92.64%) types. Another interesting phenomenon

8   was that the proportion of GC-AG sites (7.65%) in IR was higher than in the other AS

9   types (Additional file 8: Figure S6).

10 **Genomic features influence AS frequency and AS event types**

11 We grouped together the distribution of AS events, along with other characteristics of

12 the gene, across each of the 21 chromosomes in order to gain a better understanding

13 of the relationship of AS with other characteristics of the genome (Fig. 5A). A

14 previous study in soybean illustrated how the length of the intron, the number of

15 exons, the level of gene expression, and the content of GC can influence the AS

16 frequency [14]. To explore these relationships in wheat, we performed correlation

17 analysis between the number of AS events in individual genes with the length of

18 introns, the number of exons, the level of expression, and the content of GC. The

19 results demonstrated that the number of AS did not display a significant correlation

20 with the level of gene expression, but showed significant positive correlations for the

21 length of the introns and the number of exons, and displayed a significant negative

22 correlation with the content of GC (Additional file 9: Table S3). To further confirm

23 the correlations, we compared these characteristics with non-AS genes and AS genes.

24 The AS genes had significantly longer introns, more exons, and lower GC contents

25 than did the non-AS genes, however, we did not observe any significant differences in

26 the levels of expression between the AS and non-AS genes (Fig. 5B). This analysis

27 suggests that the frequency of AS in wheat displays a positive correlation with the

1 length of the intron and the number of exons, and displayed a negative correlation

2 with the content of GC.

3 We also determined the correlations between the gene features, including the length of

4 the intron, number of exons, levels of expression, and content of GC, as well as the

5 numbers of each type of AS event. The results showed that the frequency of these AS

6 events was significantly positively correlated with both the length of the intron and

7 the number of exons, and displayed a negative correlation with the content of GC

8 (Additional file 10: Table S4). This was comparable to our results for all AS events

9 (Additional file 9: Table S3). We chose genes that displayed only one of the four AS

10 types to compare, in order to understand the differences between AS types. Our

11 findings indicated that the AD, AA, and ES genes had relatively longer introns, a

12 higher number of exons, and higher levels of expression, but lower content of GC

13 than the IR types (Fig. 5C). These results indicate that different AS gene types have

14 different features, which might substantially influence the specific AS types.

15 **Changes in alternative splicing during seed development**

16 Some have indicated that AS differs significantly among tissues/organs. In this study,

17 we identified large numbers of AS genes in the developing grain samples, including

18 6,992, 7,952, 7,881, and 6,958 AS genes in the GR5, GR10, GR15, and GR20

19 samples, respectively (Additional file 6: Figure S4). From the expression data in the

20 four grain samples, we were able to gain a comprehensive view of the dynamic

21 expression changes in AS genes during grain development. Hierarchical clustering of

22 the expression of these genes shows that there are distinct patterns during

23 development (Additional file 11: Figure S7A), and further trend analysis showed that

24 there are 10 significant profiles for the expression of these AS genes (P <0.001;

25 Additional file 11: Figure S7B). To obtain an overview of the biological processes

26 involved in the AS genes that are related to seed development, we performed Gene

1 Ontology (GO) enrichment analyses. Generally, several GO categories related to

2 metabolism and development, such as "primary metabolism" (GO:0044238), "lipid

3 metabolism" (GO:0006629), "nitrogen compound metabolism" (GO:0006807),

4 "nucleobase-containing compound metabolic" (GO:0006139), and "multicellular

5 organism development" (GO:0007275), were found to be overrepresented in the up-

6 regulated profiles (profiles 21, 23, 24) (Additional file 11: Figure S7C), while GO

7 categories related to signal transduction, cell structure organization, and cell cycle,

8 such as "anatomical structure morphogenesis" (GO:0009653), "cell cycle"

9 (GO:0007049), "cell differentiation" (GO:0030154), and "signal transduction"

10 (GO:0007165), were overrepresented in the down-regulated profiles (profiles 0, 1, 3,

11 4, 9) (Additional file 11: Figure S7C).

12 To further elucidate the important functions of AS in seed development, we screened

13 the grain-specific AS genes. Ten tissue/organ samples were divided into four groups

14 based on the similarity of growth and developmental stages: seedling, flag leaf, spike,

15 and grain. A search among the differentially expressed AS dataset identified a great

16 number of tissue/organ-specific AS genes, and 2,737 seed-specific AS genes were

17 found across all tissue/organ samples (Fig. 6A). Among these seed-specific AS genes,

18 1,456, 1,342, 1,245, and 1,160 AS genes were found in the GR5, GR10, GR15, and

19 GR20 samples, respectively (Fig. 7A). The number of AS genes in the earlier seed

20 developmental stages were significantly higher than in the later developmental stages,

21 and the changes that occur mirror the phases of development. GO enrichment analysis

22 revealed that GO categories related to metabolism and development, such as "primary

23 metabolism" (GO:0044238), "glycoprotein metabolism" (GO:0009101), "RNA

24 metabolism" (GO:0016070), "starch metabolism" (GO:0005982), and "post-

25 embryonic development" (GO:0009791) are overrepresented (Fig. 6B). Wheat grain

26 development includes three main phases; (1) fertilization, which initiates grain

27 development, (2) the formation stage of the grain between 0-10 days post-anthesis

1 (DPA), in which the embryos and endosperm cells form and proliferate rapidly, and

2 (3) the grain filling stage at 11-23 DPA, in which starch and other reserves

3 accumulate and the endosperm cells enlarge rapidly. Grain development reaches the

4 mature stage at 23 DPA [43]. We further performed GO enrichment analyses on the

5 first (GR5 and GR10) or second (GR15 and GR20) phase-specific AS genes (Fig. 7B).

6 Of the first phase-specific AS genes, GO categories related to "post-embryonic

7 development" (GO:0009791), "anatomical structure development" (GO:0048856),

8 "cell cycle" (GO:0007049), and "cellular component organization or biogenesis"

9 (GO:0071840), were enriched (Fig. 7C). For the second phase-specific AS genes, the

10 enriched GO categories were related to "nucleobase-containing compound metabolic"

11 (GO:0006139), "lipid metabolism" (GO:0006629), "nitrogen compound metabolism"

12 (GO:0006807), and "primary metabolism" (GO:0044238) (Fig. 7C).

13 Trend analysis of the seed-specific AS genes revealed 20 gene expression profiles

14 (Additional file 12: Figure S8), and six of these were identified as significant profiles

15 (P <0.001; Fig. 6C, D). Of the 20 gene expression profiles, profiles 0, 6, 18, 19, and

16 20 contained much more AS genes than did the other profiles (Fig. 6C). We also

17 applied enrichment analysis to the seed-specific AS genes in the significant profiles.

18 Of the significant profiles, profile 6 had many more AS genes than the other profiles,

19 profile 19 contained the genes that were continuously up-regulated during grain

20 development, and profile 0 contained the genes that were continously down-regulated

21 (Fig. 6C, D). In the GO annotation analysis of profile 0 (down-regulated profile; P =

22 4.4E-165), "response to external stimulus" (GO:0009605), "gene silencing"

23 (GO:0016458), "regulation of gene expression" (GO:0040029), "meristem structural

24 organization (GO:0009933), "anatomical structure arrangement" (GO:0048532),

25 "microtubule-based process" (GO:0007017), "cell cycle process" (GO:0022402), and

26 "actin filament organization" (GO:0007015) were the most significantly enriched

27 terms; this suggested that the AS genes related to signal transduction, cell structure

1 organization, and cell cycle processes are active in the early phase of grain

2 development, but are gradually silenced as grain development progresses. However,

3 in profile 19 (up-regulated profile; $P = 1.5E-118$), "monocarboxylic acid metabolic

4 process" (GO:0032787), "cofactor metabolic process" (GO:0051186), "tetraterpenoid

5 metabolic process" (GO:0016108), "small molecule metabolic process"

6 (GO:0044281), and "starch metabolic process" (GO:0005982) were the most

7 significantly enriched GO terms (Additional file 13: Table S5). This is in line with the

8 process of reserve accumulation, which can provide energy for seed germination,

9 post-germination and growth, and is initiated in early grain development and ends

10 when the seed matures. Taken together, the changes in AS genes during grain

11 development corresponded to the physiological and biochemical processes of grain

12 development from the "division and expansion" to "grain filling" and finally the

13 "maturation and desiccation" stages.

DISCUSSION

15 Our analysis of RNA-seq is currently the most comprehensive study on alternatively

16 spliced transcripts in wheat, and will be extremely important when studying

17 expression regulation as well as the function of genes and proteins. We demonstrated

18 that 18,868 genes containing intron are alternatively spliced in wheat and that each AS

19 gene undergoes an average of five to six AS events (90,546 AS events/18,868 AS

20 genes) (Fig. 4A). As expected, the typical plant intron junctions GT-AG make up the

21 highest incidence of intron terminal dinucleotide signals within the AS events we

22 observed, with GT-AG introns representing the next biggest group (Additional file 8:

23 Figure S6). Unfortunately, the samples in this study were primarily obtained from

24 different tissues or the same tissues at varying stages of development, and didn't

25 include samples from plants exposed to stressful environmental conditions. Previous

26 research has found that a large number of AS events are brought on only by biotic and

27 abiotic stresses [10]. As such, AS events in wheat could be more widespread than we

1   first thought.

2   The assembling of transcripts out of short sequencing reads is a difficult process,

3   making it necessary to procure assistance for the transcriptome assembly. We used a

4   high-resolution alternative splicing RT-PCR panel to analyze a substantial dataset of

5   transcripts and AS events, attaining a high level of validation due to the experimental

6   assistance for 32 transcripts that were predicated by RNA-seq for the regions on the

7   RT-PCR panel (Additional file 5: Figure S3). Our analysis demonstrated that some of

8   the detected AS genes alternatively spliced in each sample we studied. Subsequent

9   GO analysis found that these AS genes were highly expressed in mechanisms

10   associated with metabolism, development, and RNA splicing (Fig. 3F). These findings

11   are partially backed up by the fact that genes related to RNA splicing are also

12   regulated by AS and other splicing characteristics [44, 45]. Other than these

13   constitutively alternatively spliced genes, there are many tissue/organ-specific AS

14   genes. Our study demonstrated that the total AS events, AS genes, and the frequencies

15   of the individual AS types differed significantly among the various tissues/organs (Fig.

16   3E; Fig. 4B; Additional file 6: Figure S4). The AS frequency was higher in spikes at

17   the heading stage (YS15) than in other tissues. Similar observations were made in

18   quickly growing tissues in soybean, like the meristem of the shoot or budding seeds

19   [14]. It is believed that AS is a primary factor involved in the regulation of the

20   expression of genes, in addition to increasing proteome diversity [9]. These

21   tissue/organ-specific AS events could serve in an organized fashion in particular

22   mechanisms or networks as gene groups co-regulated at the transcriptional level [46].

23   In seed development, the signal transduction-, cell structure organization-, and cell

24   cycle process-related AS genes are active early in the developing grains, but are

25   gradually silenced as the process of grain development proceeds. We also found that

26   reserve accumulation-related AS genes were up-regulated from the early to the middle

27   developmental stages of the wheat grain (Fig. 7; Additional file 11: Figure S7;

1   Additional file 13: Table S5).

2   The structure of plant and animal genes differ significantly in terms of the lengths of

3   the introns. In animals, the average relative length of the introns is large, but is much

4   smaller in plants [13]. The differences in the size of the introns between animals and

5   plants could be due to the different ways in which they process introns [47], which

6   then results in different levels of various AS types between animals and plants. In

7   animals, ES is the most commonly found AS event [19, 48,17], while in plants, IR is

8   the most commonly found [5, 11, 13, 14]. Our study demonstrated that IR is the most

9   common form of AS in wheat, which is similar to other plants. Our findings also

10   indicate that the ratios of the four different types of AS in plants could be different in

11   monocots and dicots (Fig. 4C). While IR typically accounts for most AS events in

12   plant genomes, the proportions of AS types show significant differences in various

13   species of plant (Fig. 4C). In our analysis, the ratios of both AD and AA were higher

14   than the ratio of ES (Fig. 4A), which aligns with the findings from *Arabidopsis* [5, 11,

15   13]. The differences between the ratios of the different types of AS in the four

16   monocots wheat, *B. distachyon*, rice, and maize, as well as in the four reference dicots

17   *Arabidopsis*, potato, soybean, and poplar demonstrate that the mechanisms relating to

18   the processing of the introns or AS could have diverged between monocots and dicots.

19   As the number of exons increases, the addition of splice variants is also thought to

20   increase [49]. Our results suggest that the structure of genes can have a significant

21   effect on the frequency of AS. An examination of the data from individual genes

22   showed that AS frequency displayed a positive correlation with both the length of the

23   introns and the number of exons, and displayed a negative correlation with the content

24   of GC (Fig. 5). Statistical analysis showed that AS genes possessed more exons and

25   longer introns than non-AS genes, indicating that changes in the length of the introns

26   and the number of exons could be the primary factors affecting variation in AS (Fig.

27   3). Our findings suggest that differences in the structure of the gene, particularly the

length of the introns and the number of exons, could have a meaningful effect on the frequency of AS.

CONCLUSION

In summary, through transcriptome analysis, over ten thousand AS genes were identified and characterized in wheat cultivar Xiaoyan 6. Especially 2,737 seed-specific AS genes were identified. the results of this study offer a significant amount of important AS events across the genome, making it a particularly important source of information for better understanding the regulation of genes. Further research is needed to better understand the molecular processes governing the splicing factors, their transcript targets, and the effects of isolated alternative splicing events.

MATERIALS AND METHODS

Plant growth conditions and sample collection

We used the wheat cultivar Xiaoyan 6 as experimental material in this study. We grew the wheat plants during the traditional growing season at an experimental plot located in Northwest A & F University, Yangling, Shanxi, China (longitude 108°E, latitude 34°15'N) from 2014 to 2015. Ten tissue/organ samples including roots, stems, and leaves of wheat seedlings at the five-leaf stage, young spikes at the early booting stage, spikes at the heading stage, flag leaves at the heading stage, and grain at 5, 10, 15, and 20 DPA. We collected each specimen from a minimum of five separate plants for two replicates, and the mixed samples were flash-frozen using liquid nitrogen before RNA extraction.

Isolation of RNA, construction of the cDNA library, and high-throughput sequencing

We isolated total RNA using the Total RNA Rapid Extraction Kit for Polysaccharides Polyphenol Plant (BioTeke) according to the manufacturer's directions. The

preliminary determinations of both the purity and levels of the RNA samples were performed using a NanoDrop 2000 spectrophotometer (Thermo) and RNase-free agarose gel electrophoresis, respectively. We then treated the extracted RNA with RNase-free DNase I (New England Biolabs) at 37 °C for 30 min to remove the residual DNA. We used the TruSeq paired-end mRNA-Seq kit, multiplex adapter ligation, and 125-based paired-end sequencing on the Illumina Hiseq 2500 to prepare a library from the resulting total RNA.

Read alignment and putative transcript assemblies

T. aestivum genome sequences and annotation (IWGSC RefSeq v1.0) were downloaded from Ensemble Plants (http://archive.plants.ensembl.org/Triticum_ aestivum/Info/Annotation/#assembly). We used FASTQC tools to remove sequences that were of poor quality, contained adapters, and contained nucleotide sequences that were ambiguous from the successfully prefiltered reads. This filtered for additional quality. We then used the Tuxedo RNA-seq analysis pipeline [37,38] to map the RNA-seq reads to the reference genome of *T. aestivum*. This tuxedo pipeline is made of Cuffdiff2 (differential gene and transcript expression, and differential splicing analysis), Cufflinks2 (transcript assembly), Cuffmerge2 (transcript merging), and TopHat2 (mapping). We used the Integrated Genome Viewer (Broad Institute) and Seqmonk tools (Babraham Bioinformatics) to visualize, at the chromosome-level, the read densities. These were performed according to the recommended program documentation.

We used Cufflinks 2 [37] to piece together the genome-matched reads from each RNA-seq library with two changes from the standard protocol: a maximum intron size of 100,000 and a minimum intron size of 20. We used Cuffmerge 2 [37] to merge the specific transcript groups into a single dataset of transcripts. Annotating the new junctions needed a minimum of 10 reads, and all new transcripts had to represent a

minimum of 10% of the total gene abundance in a minimum of one library. Cuffnorm [50]was used to count and analyze the new transcripts in each genotype and tissue, using default parameters. We used the FPKM mapped fragments metric [37] to normalize the RNA-seq expression.

AS event identification

We mapped the various assembled transcript isoforms to the related model of the gene via Cuffcompare, a program included in the Cufflinks program and using a procedure previously described [37], in order to detect AS events. We then used ASTALAVISTA [39] to identify the AS, analyzing four kinds of AS events: AA, ES, AD, and IR [51].

RT-PCR validation

The selected AS events were analyzed, via a set of primers, by RT-PCR (Additional file 14: Table S6) which we built for each AS event. Total RNAs from ten tissue/organ samples, including root, stem, leaf of wheat seedling at five-leaf stage, young spike at early booting stage, spike at heading stage, flag leaf at heading stage, and the grain of 5, 10, 15 and 20 DPA, were extracted using with TRIzol reagent, according to the manufacturer's instructions, and then treated with DNAase I, after which they were reverse-transcribed to cDNA (random priming).

Statistical analysis

We used Pearson's correlation with bootstrap resamplings to analyze the correlation between the characteristics we studied. The method was the same as previously described [52]. We also used an ANOVA *t*-test to compare the levels of expression and characteristics of the genome between the AS and non-AS genes as well as between the low-AS and high-AS genes. We used Fisher's LSD test to compare the various characteristics of genes with different AS events, and between different kinds of AS events.

1 **Accession numbers**

2 Raw data relating to the RNA-seq from the 20 specimens used in this study are

3 accessible from the Wheat transcriptome sequencing database (http://genedenovoweb.

4 ticp. net:81/Wheat_GDR1246/index.php).

5

6

7 **Abbreviations**

8 AA: alternative acceptor; AD: alternative donor; AS: Alternative splicing; ES: exon

9 skipping; EST: expressed sequence tag; FPKM: Fragments Per Kilobase of exon

10 model per Million mapped fragments; GO: Gene Ontology; IR: intron retention; PCA:

11 principal component analysis;PTC: premature termination codon; snRNA: small

12 nuclear RNA

17 **Author Contributions**

18 HZ and HM designed the work. HM, JZ, LG, and FM analyzed the data. LG and QC

19 performed the experiments. XL helped in seedling planting and sample preparation.

20 HM and HZ wrote the manuscript with input from JZ, LG, QC, FM, and XL. Each

21 author has studied and authorized the final version of this study.

Crop Stress Biology for Arid Areas, Northwest A & F University (CSBAAZD2016001).

Availability of data and materials

Raw data relating to the RNA-seq from the 20 specimens used in this study are accessible from the Wheat transcriptome sequencing database (http://genedenovoweb. ticp. net:81/Wheat_GDR1246/index.php).

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflicts of interest.

Author details

[1] State Key Laboratory of Crop Stress Biology for Arid Areas, College of Plant Protection, Northwest A & F University, Yangling, Shaanxi 712100, China. [2] State Key Laboratory of Crop Stress Biology for Arid Areas, College of Life Sciences, Northwest A & F University, Yangling, Shaanxi 712100, China

References

1. Lareau LF, Green RE, Bhatnagar RS, Brenner SE. The evolving roles of alternative splicing. Curr Opin Struct Biol. 2004;14(3):273-282.

2. Wahl MC, Will CL, Lührmann R. The spliceosome: Design principles of a dynamic RNP machine. Cell. 2009;136(4):701-718.

3. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, et al. Function of alternative splicing. Gene. 2005;344:1-20.

4. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA. 2003;100(1):189-192.

5. Wang BB, Brendel V. Genome wide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA . 2006;103(18):7175-7180.

6. Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, et al. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. Nucleic Acids Res. 2012;40(6):2454-2469.

7. Drechsel G, Kahles A, Kesarwani AK, Stauffer E, Behr J, Drewe P, et al. Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the *Arabidopsis* steady state transcriptome. Plant Cell. 2013;25(10):3726-3742.

8. Howard BE, Hu Q, Babaoglu AC, Chandra M, Borghi M, Tan X, et al. High-throughput RNA sequencing of pseudomonas- infected *Arabidopsis* reveals hidden transcriptome complexity and novel splice variants. PLoS One. 2013;8(10):e74183.

9. Reddy AS, Marquez Y, Kalyna M, Barta A. Complexity of the alternative splicing landscape in plants. Plant Cell. 2013;25(10):3657-3683.

1   10. Staiger D, Brown JW. Alternative splicing at the intersection of biological timing,
2   development, and stress responses. Plant Cell. 2013;25(10):3640-3656.

3   11. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Genome-
4   wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res.
5   2010;20(1):45-58.

6   12. Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, et al. Deep RNA sequencing at
7   single base-pair resolution reveals high complexity of the rice transcriptome.
8   Genome Res. 2010;20(5):646-654.

9   13. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey
10   reveals increased complexity of the alternative splicing landscape in *Arabidopsis*.
11   Genome Res. 2012;22(6):1184-1195.

12 14. Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, et al. Global dissection of
13   alternative splicing in paleopolyploid soybean. Plant Cell. 2014;26(3):996-1008.

14 15. Thatcher SR, Zhou W, Leonard A, Wang BB, Beatty M, Zastrow-Hayes G, et al.
15   Genome-wide analysis of alternative splicing in *Zea mays*: landscape and genetic
16   regulation. Plant Cell. 2014;26(9):3472-3487.

17 16. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev
18   Biochem. 2003;72:291-336.

19 17. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative
20   isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470-
21   476.

22 18. Iida K, Go M. Survey of conserved alternative splicing events of mRNAs
23   encoding SR proteins in land plants. Mol Biol Evol. 2006;23(5):1085-1094.

24 19. Modrek B, Lee C. A genomic view of alternative splicing. Nat Genet.
25   2002;30(1):13-19.

26 20. Buratti E, Baralle FE. Influence of RNA secondary structure on the pre-mRNA

splicing process. Mol Cell Biol. 2004;24(24):10505-10514.

21. Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. Epigenetics in alternative pre-mRNA splicing. Cell. 2011;144(1):16-26.

22. Schor IE, Rascovan N, Pelisch F, Alló M, Kornblihtt AR. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc Natl Acad Sci USA. 2009;106(11):4325-4330.

23. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science. 2010;327(5968):996-1000.

24. Cramer P, Pesce CG, Baralle FE, Kornblihtt AR. Functional association between promoter structure and transcript alternative splicing. Proc Natl Acad Sci USA. 1997;94(21):11456-11460.

25. Pagani F, Stuani C, Zuccato E, Kornblihtt AR, Baralle FE. Promoter architecture modulates CFTR exon 9 skipping. J Biol Chem. 2003;278(3):1511-1517.

26. Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogues G. Multiple links between transcription and splicing. RNA. 2004;10(10):1489-1498.

27. Auboeuf D, Dowhan DH, Kang YK, Larkin K, Lee JW, Berget SM, et al. Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes. Proc Natl Acad Sci USA. 2004;101(8):2270-2274.

28. Rosenfeld MG, Amara SG, Roos BA, Ong ES, Evans RM. Altered expression of the calcitonin gene associated with RNA polymorphism. Nature. 1981;290(5801):63-65.

29. Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. Genome Res. 1999;9(12):1288-1293.

30. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. Alternative splicing and genome

1   complexity. Nat Genet. 2002;30(1):29-30.

2   31. Zhu W, Brendel V. Identification, characterization and molecular phylogeny of

3   U12-dependent introns in the *Arabidopsis thaliana* genome. Nucleic Acids Res.

4   2003;31(15):4561-4572.

5   32. Zhu W, Schlueter SD, Brendel V. Refined annotation of the *Arabidopsis* genome

6   by complete expressed sequence tag mapping. Plant Physiol. 2003;132(2):469-48.

7   33. Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, et al. Genome-wide

8   analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-

9   length cDNA sequences. Nucleic Acids Res. 2004;32(17):5096-5103.

10   34. Xiao YL, Smith SR, Ishmael N, Redman JC, Kumar N, Monaghan EL, et al.

11   Analysis of the cDNAs of hypothetical genes on *Arabidopsis* chromosome 2

12   reveals numerous transcript variants. Plant Physiol. 2005;139(3):1323-1337.

13   35. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. Comprehensive

14   analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*.

15   BMC Genomics. 2006;7:327.

16   36. Langridge P. Genomics: Decoding our daily bread. Nature. 2012;491(7426):678-

17   680.

18   37. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential

19   gene and transcript expression analysis of RNA-seq experiments with TopHat and

20   Cufflinks. Nat Protoc. 2012;7(3):562-578.

21   38. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter, L.

22   Differential analysis of gene regulation at transcript resolution with RNA-seq.

23   Nat Biotechnol. 2013;31(1):46-53.

24   39. Foissac S, Sammeth M. ASTALAVISTA: dynamic and flexible analysis of

25   alternative splicing events in custom gene datasets. Nucleic Acids Res.

26   2007;35:W297-W299.

1 40. Chen TW, Wu TH, Ng WV, Lin WC. Interrogation of alternative splicing events
2   in duplicated genes during evolution. BMC Genomics. 2011;12:S16.

3 41. Thanaraj TA, Clark F. Human GC-AG alternative intron isoforms with weak
4   donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids
5   Res. 2001;29(12):2581-2593.

6 42. Churbanov A, Winters-Hilt S, Koonin EV, Rogozin IB. Accumulation of GC
7   donor splice signals in mammals. Biol Direct. 2008;3:30.

8 43. Johnson DR, Tanner JW. Calculation of the rate and duration of grain filling in
9   corn (*Zea mays* L.). Crop Sci. 1972;12:485-486.

10 44. Saltzman AL, Pan Q, Blencowe BJ. Regulation of alternative splicing by the core
11   spliceosomal machinery. Genes Dev. 2011;25(4):373-384.

12 45. Thomas J, Palusa SG, Prasad KV, Ali GS, Surabhi GK, Ben-Hur A, et al.
13   Identification of an intronic splicing regulatory element involved in auto-
14   regulation of alternative splicing of *SCL33* pre-mRNA. Plant J. 2012;72(6):935-
15   946.

16 46. Blencowe BJ. Alternative splicing: New insights from global analyses. Cell.
17   2006;126(1):37-47.

18 47. Hartmuth K, Barta A. In vitro processing of a plant pre-mRNA in a HeLa cell
19   nuclear extract. Nucleic Acids Res. 1986;14(19):7513-7528.

20 48. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A
21   global view of gene activity and alternative splicing by deep sequencing of the
22   human transcriptome. Science. 2008;321(5891):956-960.

23 49. Kopelman NM, Lancet D, Yanai I. Alternative splicing and gene duplication are
24   inversely correlated evolutionary mechanisms. Nat Genet. 2005;37(6):588-589.

25 50. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts
26   in annotated genomes using RNA-Seq. Bioinformatics. 2011;27(17):2325-2329.

1 51. Sammeth M, Foissac S, Guigó R. A general definition and nomenclature for

2   alternative splicing events. PLoS Comput Biol. 2008;4(8):e1000147.

3 52. Zhang L, Gaut BS. Does recombination shape the distribution and evolution of

4   tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? Genome

5   Res. 2003;13(12):2533-2540.

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

1　**Figure Legends**

2　**Fig. 1 Summary of RNA-seq read alignment and gene identification in this study.**

3　**a** Table showing the number and percentagae of reads that aligned to the *T. aestivum*

4　genome. **b** The numbers of aligned reads and gene lengths across each chromosome.

5　The read numbers are shown by blue histograms, and the red line indicates the sum of

6　the gene lengths for each chromosome. **c** The distribution of reads based on the wheat

7　genome annotation. **d** Distribution of the relative coverage of individual genes by

8　mapped reads. **e** Overlap of computationally predicted genes and transcripts from

9　wheat cultivar Xiaoyan 6 compared with the public *T. aestivum* genome annotation

10　(IWGSC RefSeq v1.0).

11　**Fig. 2 Global analysis of transcript expression in all 20 RNA samples from wheat**

12　**cultivar Xiaoyan 6. a** Expression density plots. **b** Expression box plots. **c** Heatmap

13　showing relative gene expression. **d** Principal component analysis. Together, these

14　analyses attest to the robustness of the *T. aestivum* transcriptome analysis, showing

15　little systematic bias between the pairs of repeated samples, and also enabling

16　identification of the significant differences present among the different tissues or the

17　same tissues sampled at different developmental stages at the genomic scale.

18　**Fig. 3 Characteristics of AS genes and GO analysis of ubiquitously alternatively**

19　**spliced genes. a-d** Comparisons between AS and non-AS genes. **a** Average domain

20　number. **b** Average gene length. **c** Average exon length. **d** Average exon number. **e**

21　Distribution of AS genes among all 10 tissues/organs included in the study. **f** GO

22　biological processes enriched ($P < 0.01$) among AS genes that were alternatively

23　spliced in all of the wheat RNA samples examined.

24　**Fig. 4 Statistics and distribution of the different AS types in the *T. aestivum***

25　**genome. a** Statistics of the different AS types: IR, intron retention; AA, alternative

26　acceptor sites; AD, alternative donor sites; ES, exon skipping. **b** Numbers of

1 individual AS types in the 10 different RNA samples. **c** Relative frequencies of the

2 different AS types in four monocots (*T. aestivum*, *B. distachyon*, rice, and maize) and

3 four dicots (*Arabidopsis*, potato, soybean, and poplar).

4 **Fig. 5 Comparisons of gene features with different AS events. a** AS, transcriptome

5 profiling, and the distribution of genomic features across the wheat genome. From

6 outside to inside, the circles represent the chromosome structure, gene location, AS

7 gene location, gene expression, gene length, and GC content. **b** Comparisons of four

8 gene features between AS and non-AS genes; intron length, exon number, gene

9 expression level, and GC content. **c** Comparisons of the gene features between the

10 different AS types. Intron length, exon number, gene expression level, and GC content

11 were compared for AS types IR, AA, AD, and ES . The bars indicate the SE.

12 **Fig. 6 Functional enrichment and expression patterns of seed-specific AS genes. a**

13 Venn diagrams showing the distribution of the AS genes in four stages of

14 development, including seedlings, flag leaves, spikes ,and seeds. RSL, roots (R) stems

15 (S) and leaves (L) of seedlings at the five-leaf stage; FL, flag leaves of plants at the

16 heading stage; YS, young spikes from plants at the heading stage; GR, seed samples

17 from spikes at 5, 10, 15, and 20 days post-anthesis. **b** GO biological processes

18 enriched ($P < 0.01$) among AS genes that were specifically alternatively spliced in

19 seeds during development. **c** Expression patterns of seed-specific AS genes. These AS

20 genes can be devided into 20 expression profiles (profile 0-20) according to heatmap.

21 **d** Expression profiles that were significantly enriched for seed-specific AS genes ($P <$

22 0.01).

23 **Fig. 7 Venn diagrams and GO analysis of seed-specfic AS genes in four seed**

24 **development periods. a** Venn diagrams showing the distribution of the AS genes in

25 four seed development periods. **b** Venn diagrams of first (GR5 and GR10) or second

26 (GR15 and GR20) phase specific AS genes**. c** GO enrichment analyses of first (GR5

1 and GR10) or second (GR15 and GR20) phase specific AS genes. GR5-GR20, seed

2 samples from 5, 10, 15 and 20 days post-anthesis spikes.

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

1 **Additional file 1: Table S1. Statistics of aligned reads in each sample.**

2 **Additional file 2: Figure S1. Log2 scale of median read density in windows of 1**

3 **kb for the 21 *T. aestivum* chromosomes.**

4 **Additional file 3: Table S2. Information about the novel genes detected in this**

5 **study.**

6 **Additional file 4: Figure S2. Pairwise expression scatter matrix for each pair of**

7 **repeated samples.**

8 **Additional file 5: Figure S3. Validation of splice variants by RT-PCR.** The primer

9 sequences used to validation of AS are listed in the Additional file 14: Table S6.

10 **Additional file 6: Figure S4. AS gene distributions and their relationships with**

11 **the expressed gene number and expression level in different tissues. a** AS gene

12 number in different tissues. **b** Plot of the AS gene number against the expressed gene

13 number in different tissues. **c** AS gene ratio in different tissues. **d** Plot of AS ratio

14 against gene expression level in different tissues.

15 **Additional file 7: Figure S5. Statistics and distribution of the different AS types**

16 **among all tissues examined.** IR, intron retention; AA, alternative acceptor sites; AD,

17 alternative donor sites; ES, exon skipping.

18 **Additional file 8: Figure S6. Splice site usage of the different types of AS events.**

19 Proportions of splice site usage of all AS events (All), intron retention (IR),

20 alternative acceptor sites (AA), alternative donor sites (AD), and exon skipping (ES)

21 are shown.

22 **Additional file 9: Table S3. Correlations of AS number with other features for**

23 **individual genes.**

24 **Additional file 10: Table S4. Correlations of AS number with gene features for**

1   **individual AS type.**

2   **Additional file 11: Figure S7. Expression patterns (a), trends analysis (b), and**

3   **GO enrichment (c) of AS genes during seed development.** GR5-GR20, seed

4   samples from 5, 10, 15 and 20 days post-anthesis spikes.

5   **Additional file 12: Figure S8. Expression profiles of seed-specific AS genes.**

6   Different colour indicates the significantly enriched profiles ($P < 0.01$).

7   **Additional file 13: Table S5. GO annotation of significantly enriched profiles of**

8   **seed-specific AS genes ($P < 0.01$).**

9   **Additional file 14: Table S6. Gene primers used for PCR and sequencing**
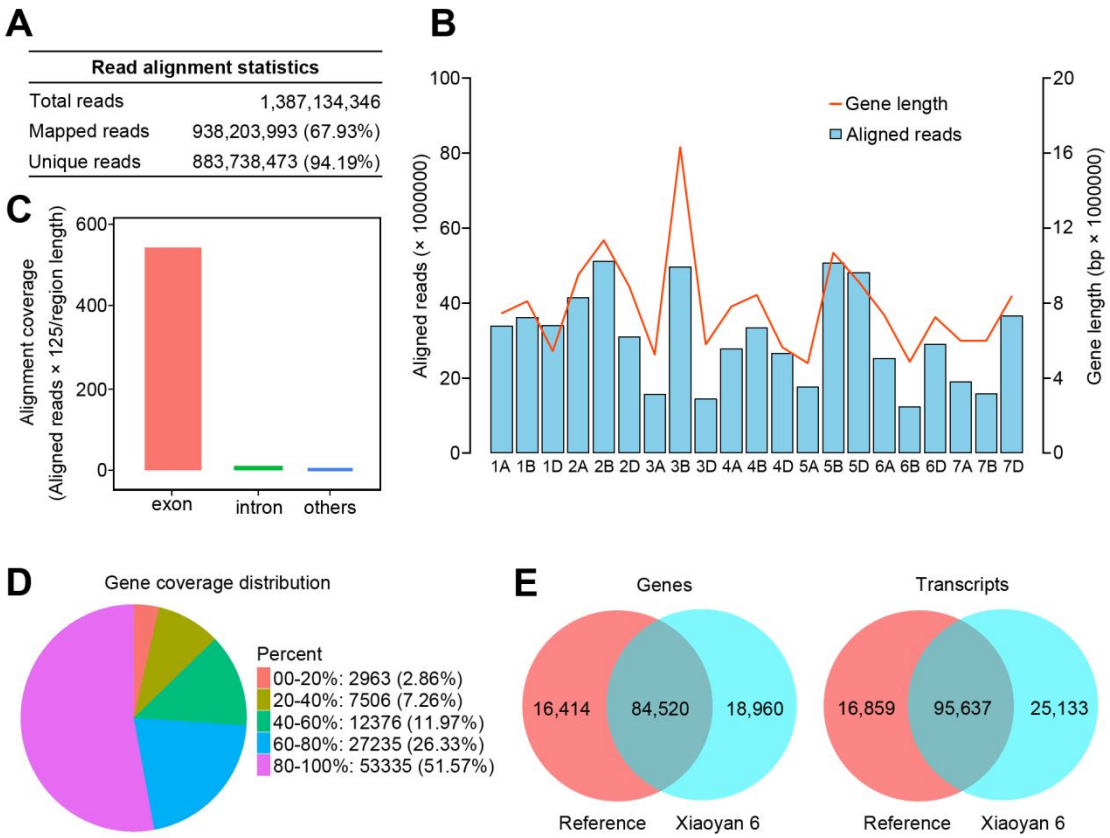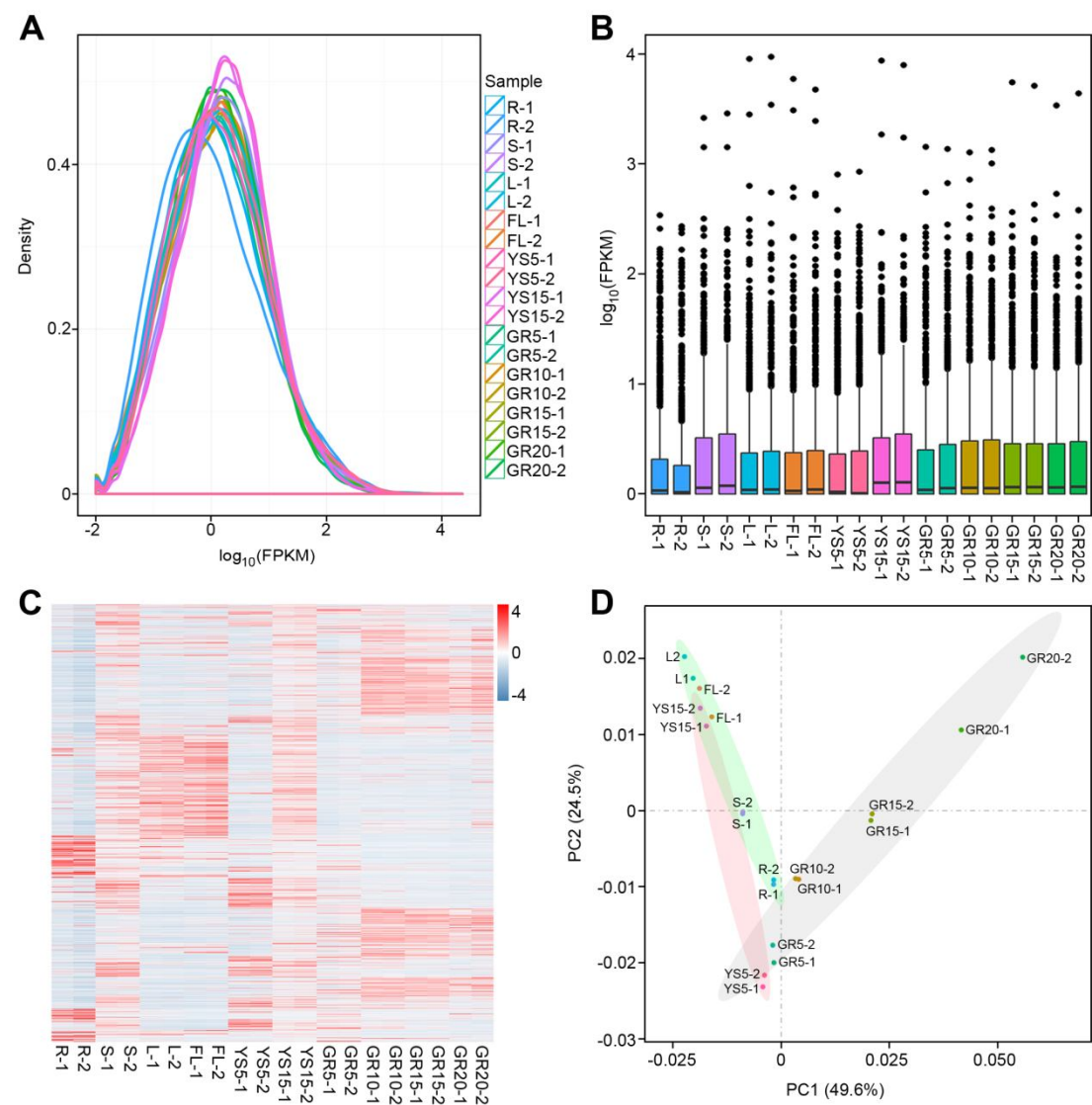
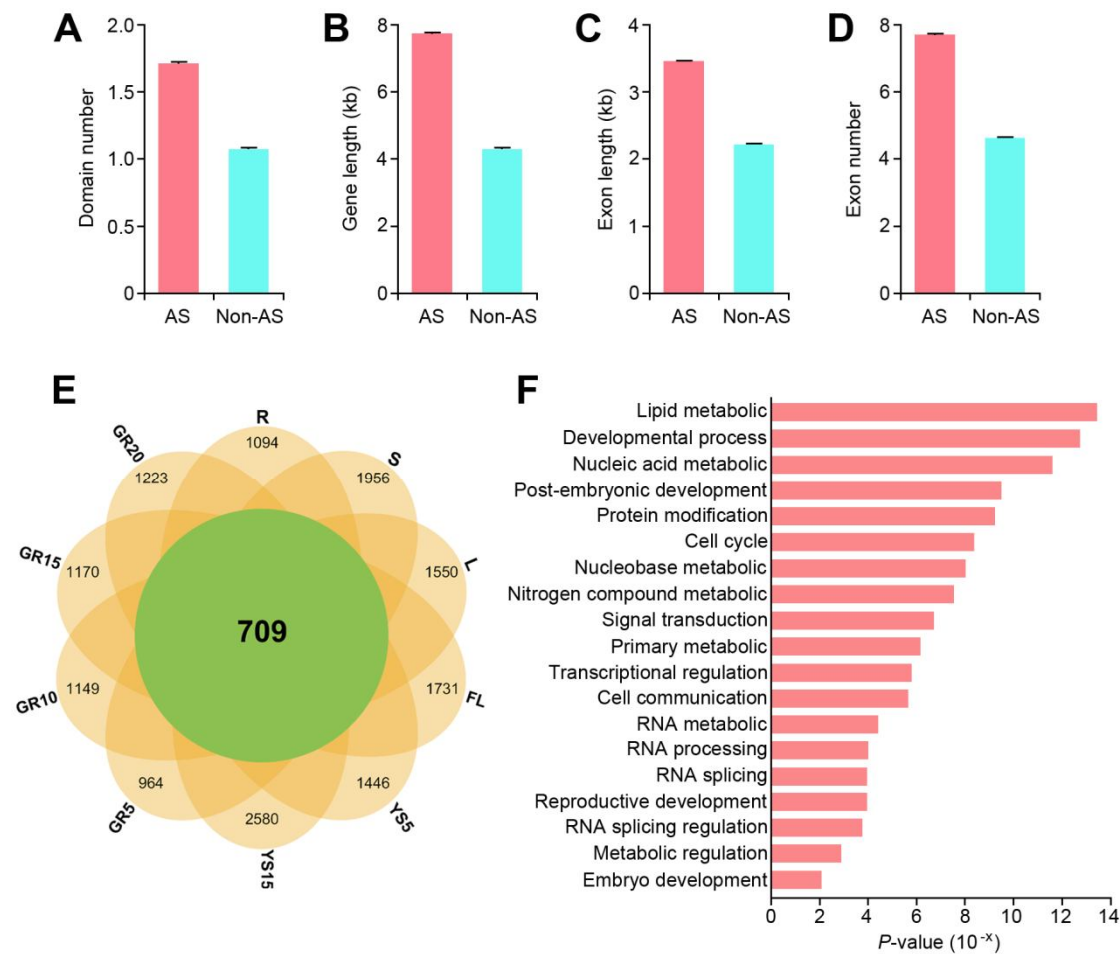10   **analyses.**

11

12

13

14

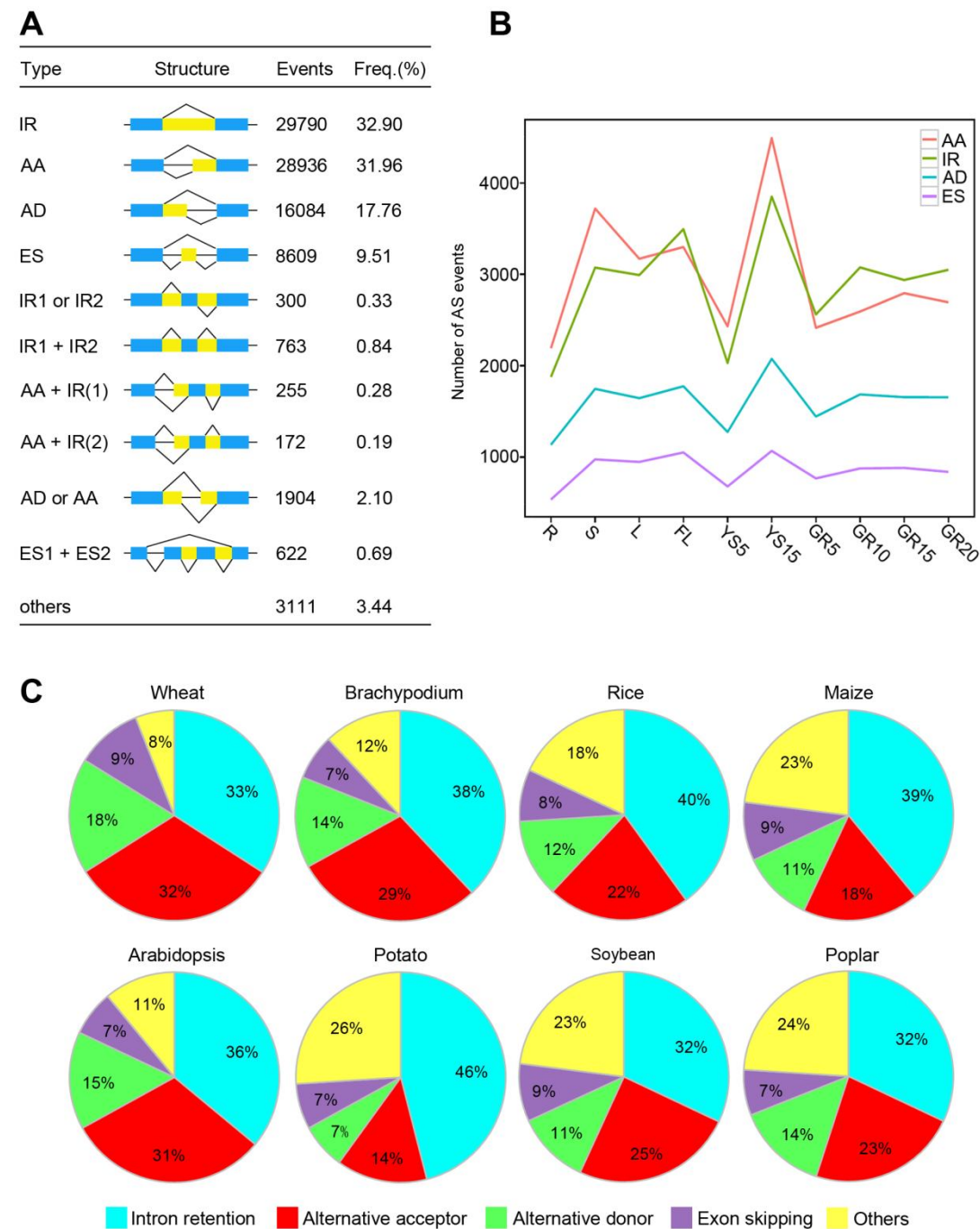**Figure 1**

Figure 2

Figure 3

Figure 4

A

Type	Structure	Events	Freq.(%)
IR		29790	32.90
AA		28936	31.96
AD		16084	17.76
ES		8609	9.51
IR1 or IR2		300	0.33
IR1 + IR2		763	0.84
AA + IR(1)		255	0.28
AA + IR(2)		172	0.19
AD or AA		1904	2.10
ES1 + ES2		622	0.69
others		3111	3.44

B

Number of AS events

AA
IR
AD
ES

R S L FL YS5 YS15 GR5 GR10 GR15 GR20

C

Wheat: 33%, 32%, 18%, 9%, 8%
Brachypodium: 38%, 29%, 14%, 7%, 12%
Rice: 40%, 22%, 12%, 8%, 18%
Maize: 39%, 18%, 11%, 9%, 23%
Arabidopsis: 36%, 31%, 15%, 7%, 11%
Potato: 46%, 14%, 7%, 7%, 26%
Soybean: 32%, 25%, 11%, 9%, 23%
Poplar: 32%, 23%, 14%, 7%, 24%

■ Intron retention ■ Alternative acceptor ■ Alternative donor ■ Exon skipping ■ Others

36

Figure 5

Figure 6

A

B

| Regulation of metabolism |
| Transcriptional regulation |
| Primary metabolism |
| Developmental process |
| Cell cycle |
| Glycoprotein metabolism |
| Nitrogen compound metabolism |
| Reproductive process |
| RNA metabolism |
| Nucleic acid metabolism |
| Cytoskeleton organization |
| Cellular component organization |
| Post-embryonic development |
| Lipid metabolism |
| Protein phosphorylation |
| RNA processing |
| Response to stimulus |
| RNA metabolism |
| Histone modification |
| Gene silencing |
| Carbohydrate metabolism |
| Starch metabolism |

P-value (10^{-x})

C

D

Profile 6: $P = 0$

Profile 18: $P = 8.2e\text{-}277$

Profile 19: $P = 1.5e\text{-}118$

Profile 0: $P = 4.4e\text{-}165$

Profile 3: $P = 8.2e\text{-}25$

Profile 2: $P = 0.00014$

38

Figure 7

Figures

Figure 1

A

Read alignment statistics

Total reads	1,387,134,346
Mapped reads	938,203,993 (67.93%)
Unique reads	883,738,473 (94.19%)

B

C

D Gene coverage distribution

Percent
00-20%: 2963 (2.86%)
20-40%: 7506 (7.26%)
40-60%: 12376 (11.97%)
60-80%: 27235 (26.33%)
80-100%: 53335 (51.57%)

E

Genes

16,414 | 84,520 | 18,960

Reference | Xiaoyan 6

Transcripts

16,859 | 95,637 | 25,133

Reference | Xiaoyan 6

33

Figure 1

Summary of RNA-seq read alignment and gene identification in this study. a Table showing the number and percentagae of reads that aligned to the T. aestivum genome. b The numbers of aligned reads and gene lengths across each chromosome. The read numbers are shown by blue histograms, and the red

line indicates the sum of the gene lengths for each chromosome. c The distribution of reads based on the wheat genome annotation. d Distribution of the relative coverage of individual genes by mapped reads. e Overlap of computationally predicted genes and transcripts from wheat cultivar Xiaoyan 6 compared with the public T. aestivum genome annotation (IWGSC RefSeq v1.0).
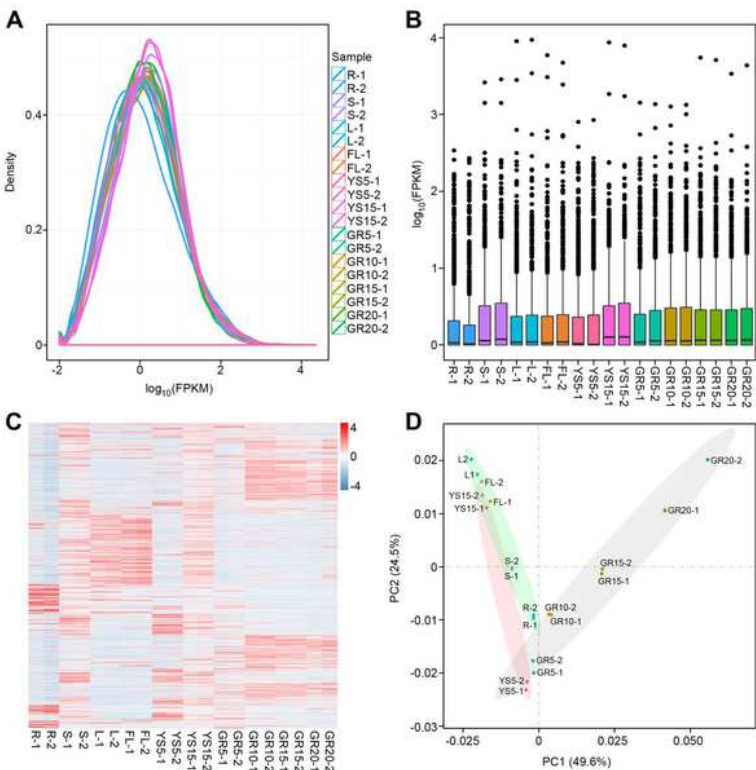

Figure 2

Figure 2

Global analysis of transcript expression in all 20 RNA samples from wheat cultivar Xiaoyan 6. a Expression density plots. b Expression box plots. c Heatmap showing relative gene expression. d Principal component analysis. Together, these analyses attest to the robustness of the T. aestivum transcriptome analysis, showing little systematic bias between the pairs of repeated samples, and also enabling identification of the significant differences present among the different tissues or the same tissues sampled at different developmental stages at the genomic scale.

Figure 3

35

Figure 3

Characteristics of AS genes and GO analysis of ubiquitously alternatively spliced genes. a-d Comparisons between AS and non-AS genes. a Average domain number. b Average gene length. c Average exon length. d Average exon number. e Distribution of AS genes among all 10 tissues/organs included in the study. f GO biological processes enriched (P < 0.01) among AS genes that were alternatively spliced in all of the wheat RNA samples examined.
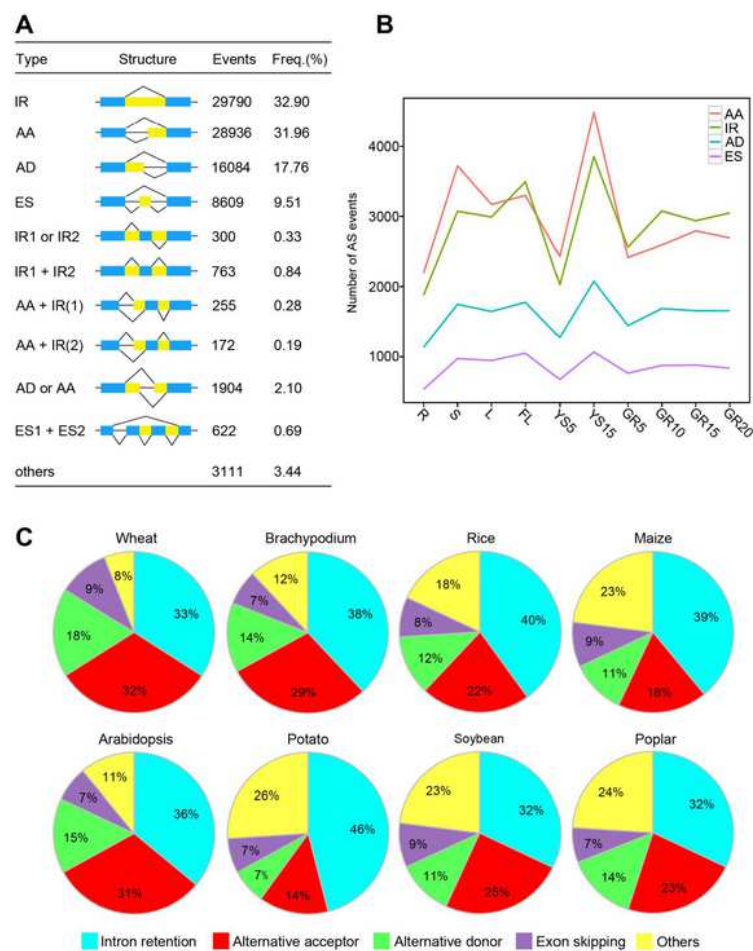

Figure 4

Figure 4

Statistics and distribution of the different AS types in the T. aestivum genome. a Statistics of the different AS types: IR, intron retention; AA, alternative acceptor sites; AD, alternative donor sites; ES, exon skipping. b Numbers of individual AS types in the 10 different RNA samples. c Relative frequencies of the different AS types in four monocots (T. aestivum, B. distachyon, rice, and maize) and four dicots (Arabidopsis, potato, soybean, and poplar).
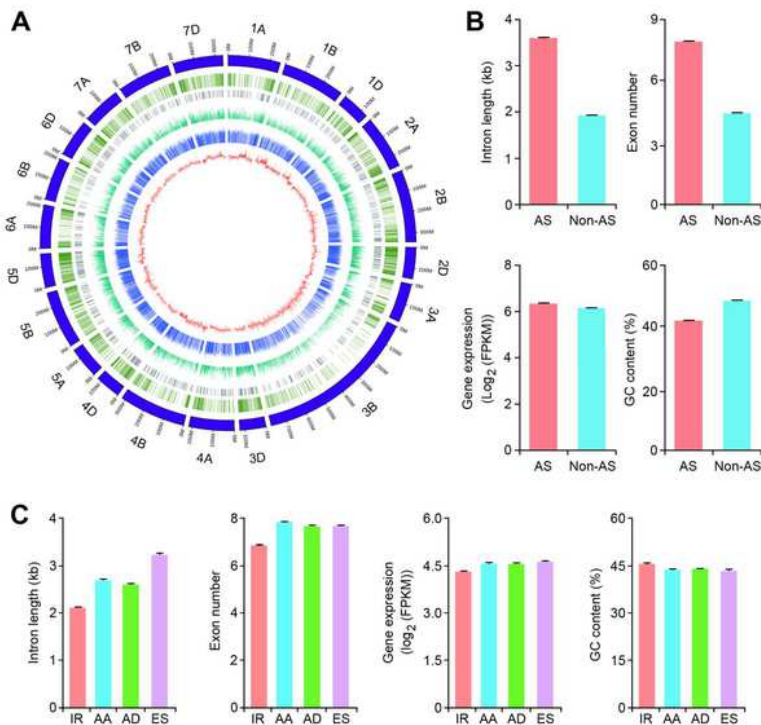

Figure 5

Figure 5

Comparisons of gene features with different AS events. a AS, transcriptome profiling, and the distribution of genomic features across the wheat genome. From outside to inside, the circles represent the chromosome structure, gene location, AS gene location, gene expression, gene length, and GC content. b Comparisons of four gene features between AS and non-AS genes; intron length, exon number, gene expression level, and GC content. c Comparisons of the gene features between the different AS types. Intron length, exon number, gene expression level, and GC content were compared for AS types IR, AA, AD, and ES . The bars indicate the SE.
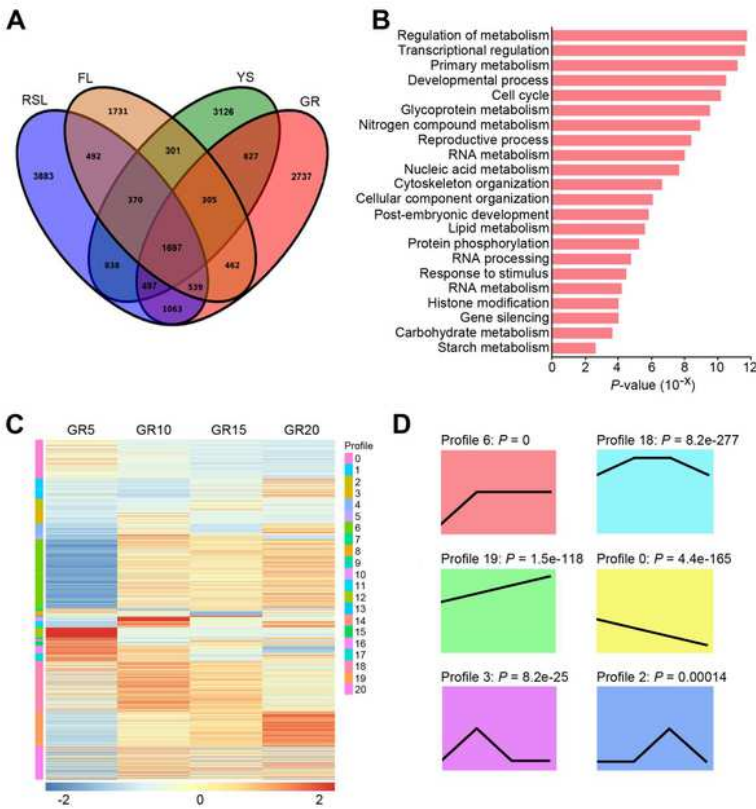
Figure 6

Figure 6

Functional enrichment and expression patterns of seed-specific AS genes. a Venn diagrams showing the distribution of the AS genes in four stages of development, including seedlings, flag leaves, spikes ,and seeds. RSL, roots (R) stems (S) and leaves (L) of seedlings at the five-leaf stage; FL, flag leaves of plants at the heading stage; YS, young spikes from plants at the heading stage; GR, seed samples from spikes at 5, 10, 15, and 20 days post-anthesis. b GO biological processes enriched (P < 0.01) among AS genes that were specifically alternatively spliced in seeds during development. c Expression patterns of seed-specific AS genes. These AS genes can be devided into 20 expression profiles (profile 0-20) according to heatmap. d Expression profiles that were significantly enriched for seed-specific AS genes (P < 0.01).
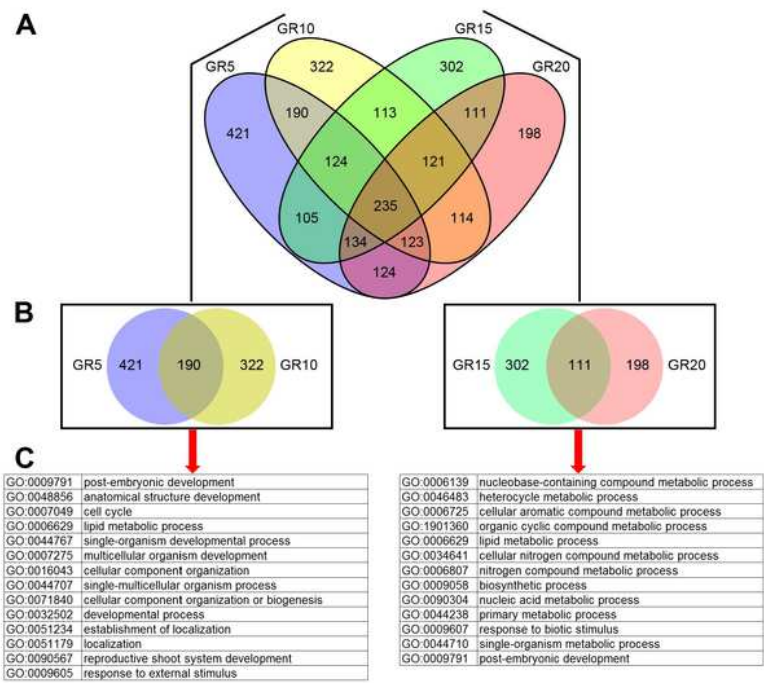
Figure 7

Figure 7

Venn diagrams and GO analysis of seed-specfic AS genes in four seed development periods. a Venn diagrams showing the distribution of the AS genes in four seed development periods. b Venn diagrams of first (GR5 and GR10) or second (GR15 and GR20) phase specific AS genes. c GO enrichment analyses of first (GR5 and GR10) or second (GR15 and GR20) phase specific AS genes. GR5-GR20, seed samples from 5, 10, 15 and 20 days post-anthesis spikes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile7FigureS5.tif
- Additionalfile12FigureS8.tif
- Additionalfile5FigureS3.tif
- Additionalfile11FigureS7.tif
- Additionalfile6FigureS4.tif
- Additionalfile4FigureS2.tif
- Additionalfile2FigureS1.tif
- Additionalfile3TableS2.xlsx
- Additionalfile8FigureS6.tif
- Additionalfile1TableS1.xls
- Additionalfile13TableS5.xlsx
- Additionalfile14TableS6.xlsx
- Additionalfile10TableS4.xlsx
- Additionalfile9TableS3.xlsx