Front. Agr. Sci. Eng. 2014, 1(3): 214–222
DOI: 10.15302/J-FASE-2014024

RESEARCH ARTICLE

Transcriptome analysis of wheat grain using RNA-Seq

Liu WEI [1*], Zhihui WU [1*], Yufeng ZHANG [1*], Dandan GUO[1], Yuzhou XU[2], Weixia CHEN[1], Haiying ZHOU[1], Mingshan YOU[2], Baoyun LI (✉)[1]

1 Beijing Key Laboratory of Crop Genetic Improvement,College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China

2 Key Laboratory of Crop Heterosis & Utilization, Ministry of Education,College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China

Abstract With the increase in consumer demand, wheat grain quality improvement has become a focus in China and worldwide. Transcriptome analysis is a powerful approach to research grain traits and elucidate their genetic regulation. In this study, two cDNA libraries from the developing grain and leaf-stem components of bread wheat cultivar, Nongda211, were sequenced using Roche/454 technology. There were 1061274 and 1516564 clean reads generated from grain and leaf-stem, respectively. A total of 61393 high-quality unigenes were obtained with an average length of 1456 bp after *de novo* assembly. The analysis of the 61393 unigenes involved in the biological processes of the grain showed that there were 7355 differentially expressed genes upregulated in the grain library. Gene ontology enrichment and the Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis showed that many transcription products and transcription factors associated with carbohydrate and protein metabolism were abundantly expressed in the grain. These results contribute to excavate genes associated with wheat quality and further study how they interact.

Keywords transcriptome analysis, wheat grain, differentially expressed genes, enrichment analysis

1 Introduction

Wheat is one of the most important crops in the world, occupying 17% of the arable land acreage worldwide, feeding about 40% of the world population and providing 20% of the total food calories and protein in human nutrition [1].

Wheat grain development is a complex process that involves coordination among a number of different tissues [2]. Most highly expressed genes in the wheat endosperm are related to storage proteins, starch metabolism enzymes and defense proteins [3]. Seed specificity of gene expression is affected by the precise control of time and space, involving interaction between cis-elements and transcription factors. Wheat storage proteins are predominantly synthesized and stably accumulated in mature endosperm tissue. The main gluten components, prolamins, are encoded by multigene families whose expression is regulated temporally and spatially. This regulation is determined by promoter elements such as the prolamin box, which comprises endosperm and nitrogen elements [4]. Transcription factors that bind to these sites include endosperm box factor-I, endosperm box factor-II and storage protein activator (SPA) in wheat grain [5]. Starch is synthesized through the coordinated interactions of several biosynthetic starch enzymes, including ADP glucose pyrophosphorylase, starch synthase, starch branching enzyme and starch debranching enzyme [6]. Therefore, understanding the molecular mechanisms by which these metabolic pathways are regulated is expected to provide valuable information for improving wheat quality.

The availability of more than one million expressed sequence tags has made it possible to use a variety of high-throughput methods to identify genes that are differentially expressed during grain development [7,8]. Moreover, transcriptome sequencing offers a manageable approach to study the complex structure and function of the wheat transcripts and can help in unraveling the genetics of wheat endosperm.

A total of 17949 publicly available expressed sequence tags generated from endosperm specific cDNA libraries have been analyzed and 2237 differentially expressed genes at six different time points (3, 7, 14, 21, 28 and 35 days post anthesis; dpa) were identified. This has elucidated much information about storage proteins,

Received August 20, 2014; accepted October 8, 2014

Correspondence: baoyunli@cau.edu.cn

*These authors contributed equally to the work

© The Author(s) 2014. This article is published with open access at http://engineering.cae.cn

putative defense proteins and proteins involved in starch and sucrose metabolism [3]. The relative expression abundances of different gene categories in the developing wheat caryopsis from 8 to 40 dpa have been analyzed using the Serial Analysis of Gene Expression method [9]. Wan studied the transcriptome expression profile of wheat cultivar, Hereward, from 6 to 42 dpa using the Affymetrix wheat GeneChip oligonucleotide arrays and found that during this period, 14550 genes were differentially expressed [10]. Yang analyzed 201 seed-specific unigenes using digital differential display technology combined with wheat chip expression and carried out the classification and gene ontology (GO) enrichment, revealing that the seed-specific unigenes participated mainly in defense reactions, stress, nutrition storage activity, enzyme inhibition and redox (reduction–oxidation) activity [11]. Xu sequenced rice, a Gramineae model plant, during different periods of embryonic development (3–5, 7 and 14 dpa) using the Illumina/Solexa sequencing technology and found that a large number of genes related to regulation of metabolism were induced by DNA replication and processing [12]. Examples include signal transduction genes mainly expressed during the early and middle embryonic development periods, protein accumulation-related genes mainly expressed during the middle embryonic development period, and starch/sucrose metabolism-related genes and protein modification-related genes expressed during the later period of embryonic development.

The Roche/454 GS FLX sequencer provides a useful tool for studying the transcriptome of wheat seed, which uses the pyrophosphate molecule released on nucleotide incorporation by DNA polymerase to fuel a downstream set of reactions that ultimately produces light from the cleavage of oxyluciferin by luciferase [13]. Roche/454 GS FLX sequencer produces an average read length of 400 bp per sample (per bead), with a combined throughput of 100 Mb of sequence data per 7h run [14,15].

Here, the Roche/454 sequencing technology was employed to study the gene expression profiles of bread wheat grain and leaf-stem during grain filling. Both the grain library and leaf-stem library were mixed to assemble as the reference sequences. The leaf-stem served as a control to the grain to obtain a profile of differentially expressed genes in the grain, specifically for mining genes associated with wheat quality.

2 Materials and methods

2.1 Plant materials

The bread wheat cultivar, Nongda211, was grown under natural conditions in Beijing. Plants were tagged at anthesis and sampled at 5-day intervals from 5 to 25 dpa and 28 dpa. The grain, flag leaf and stem below the spike were cut off and quickly frozen with liquid nitrogen separately. Frozen tissues were stored at −80°C.

2.2 RNA extraction, cDNA synthesis and sequencing

For RNA extraction, equal amounts of flag leaf and stem below the spike tissues (leaf-stem library) from each stage were pooled. Equal amounts of developing grain (grain library) were also used. Frozen tissue was ground with a mortar and pestle and approximately 100 mg of powdered tissue was sampled. RNA was isolated using the RNA extraction kit (Beijing Autolab Biotechnology Co., Ltd) according to the manufacturer's instructions. Total RNA was treated with DNase (TURBO DNase; Ambion, Austin, TX, USA) to remove DNA. RNA concentration was measured using a Nanodrop 2000 spectrophotometer (Thermo Scientific, Wilmington, USA). RNA integrity was assessed by 1% agarose gel electrophoresis.

High-quality cDNA was obtained using the SMART PCR cDNA synthesis protocol (Clontech) and checked on a 2% agarose gel to verify cDNA quality and fragment length. The cDNAs distributed between 500 and 5000 bp were collected and polyT was removed by restriction enzyme digestion (Beijing Autolab Biotechnology Co., Ltd). Approximately 3 μg of each cDNA sample was sheared into small fragments via nebulization and sequenced in a single 454 run using a GS-FLX sequencer (Beijing Autolab Biotechnology Co., Ltd).

2.3 *De novo* assembly

The raw reads were filtered to obtain high-quality clean reads prior to assembly, which was performed by removing the adaptor sequences, short sequences (less than 50 bp), contaminated sequences and low-quality sequences containing more than 20% nucleotides in a read with a *Q*-value ⩽10. Clean reads were *de novo* assembled to generate non-redundant unigenes using the Newbler assembly programs with the default parameters. The non-redundant unigenes were used for further analysis in this study.

2.4 Functional annotation

Sequence-based and domain-based alignments were used to compare sequences. Sequence-based alignments were performed against three public databases [the NCBI non-redundant database (NR), the NCBI nucleotide sequences database (NT) and the Swiss-Prot protein database) with NCBI BLAST 2.2.28 + software (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast +/2.2.28/)], using an *E*-value threshold of 10^{-5} [16]. Domain-based alignments were performed against the euKaryotic Orthologous Groups database (KOG; http://www.ncbi.nlm.nih.gov/COG/) with an *E*-value threshold of 10^{-3}. Protein structure domains of all unigenes were predicted with the HMMER 3.0 package

and hmmscan programs against the Pfam database (http://pfam.sanger.ac.uk/) [17]. Functional categorization by GO terms was performed based on the best hits from the non-redundant and Pfam databases using BLAST2GO software [18], with an *E*-value threshold of 10^{-5}. To produce the pathway annotation and identify the BRITE functional hierarchies, sequences were submitted to the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (KAAS) [19]. KAAS annotated each submitted sequence with KEGG orthology (KO) identifiers, which represent an orthologous group of genes directly linked to an object in the KEGG pathways and BRITE functional hierarchy.

2.5 Analysis of differentially expression unigenes

To analyze unigene transcript levels, the uniquely mapped reads for a specific unigene were counted by mapping reads of each library to *de novo* assembled non-redundant unigenes using Bowtie and RSEM software [20]. And the reads per kb per million mapped sequence reads (RPKM) values were computed as recommended by Mortazavi [21]. After computing the RPKM, the read count of each unigene was obtained. Differences in unigene transcript abundance between the grain and leaf-stem libraries were obtained from normalized read count values using TMM and DEGseq software [22,23]. Fold changes for each unigene between the grain and leaf-stem components were computed as the ratio of the read count values. If the read count value of the grain or the leaf-stem was 0, 0.001 was used instead of 0 to measure the fold change. The significance of differential transcript abundance was calculated using the *Q*-value (Benjamini-Hochberg corrected *P*-value) to justify the *P*-value, and only unigenes with a $|\log_2(\mathrm{FC})| \geqslant 1$ and a $Q\text{-value} \leqslant 0.005$ were used for subsequent analysis. The formula to determine the *P*-value between the grain and the leaf-stem was:

$$P(y/x) = (N_2/N_1)^y[(x+y)!]/[x!y!(1+N_2/N_1)^{(x+y+1)}]$$

where N_1 and N_2 represent the total number of clean reads mapped to all unigenes in each sample, and x and y represent the number of clean reads mapped to a common unigene in the leaf-stem and the grain libraries, respectively.

2.6 Enrichment analysis of DEGs

To assign putative biological functions and pathway involvement to the differentially expression unigenes, enrichment analysis was carried out. The GO enrichment was analyzed with the GOseq method and KEGG pathway enrichment was analyzed with KOBAS (2.0) [24,25], using Wallinius non-central hyper-geometric and hyper-geometric tests for statistical analysis, respectively. At first, all unigenes showing significant differences in transcript abundance (differentially expressed unigenes) between two libraries were mapped to the GO and KEGG pathway databases and then the number of unigenes for every GO term and KO term was calculated. The hyper-geometric test was applied to identify significantly enriched GO and KO terms from the set of DEGs. The formula for the gene enrichment test was

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where N represents the total number of unigenes with GO and KEGG pathway annotation; n represents the number of DEGs in N; M represents the number of unigenes that were annotated to certain GO or KO terms; and m represents the number of DEGs in M. The initially obtained *P*-values were then adjusted using a Benjamini—Hochberg correction and a corrected *P*-value of 0.05 was adopted as a threshold.

2.7 Unigenes verified via qPCR

Grain of each stage was ground with a mortar and pestle and approximately 100 mg of powdered tissue was sampled. RNA was isolated using the RNAprep pure plant kit (TIANGEN Biotech, China), cDNA synthesized from total RNA using the PrimeScript RT reagent kit (TAKARA, Japan) and qPCR performed using the SYBR Premix Ex Taq (TAKARA, Japan) according to the manufacturer's instructions. Primer sequences of unigenes are listed in the Appendix A (Table S1). Real-time PCR was conducted in a 7500 and 7500 Fast Sequence Detection System (Applied Biosystems). The pooled grain served as control. In each run, samples were assayed for both the target genes and ubi gene (endogenous control). Relative expression was calculated using $2^{-\Delta\Delta C_T}$ method [26].

3 Results

3.1 Transcriptome sequencing and *de novo* assembly

The bread wheat grain cDNA library and leaf-stem cDNA library were sequenced using a Roche/454 GS-FLX sequencer. Approximately 1.07 and 1.52 G clean reads were generated from the grain and the leaf-stem libraries, respectively. Then the clean reads of the two samples were mixed for assembly as reference sequences. Transcripts shorter than 100 bp were hard to annotate accurately, so only those greater than 100 bp were used in further analysis. A total of 61393 unigenes with an N50 (50% of

the assembled bases were incorporated into contigs of 728bp or longer) of 1814 and an N90 of 750 were obtained. The unigenes had an average length of 1456 bp, a minimum length of 102 bp and a maximum length of 12033 bp (Table 1).

3.2 Functional annotation of assembled unigenes

A total of 61393 distinct unigenes were acquired after assembly. Initially, all unigenes were blasted against the NCBI NR, the NCBI NT, the Swiss-Prot, the Pfam and the KOG databases. According to the hits retrieved from the Nr and Pfam databases, GO annotation was conducted. Finally, the KAAS was used to perform KEGG pathway annotation against the KEGG database. Among the 61393 unigenes, 58724 (95.7%) were returned as matches in at least one database and 7883 (12.8%) returned at least one hit in all databases (Table 2).

3.3 Global analysis of gene expression

As mentioned above, the clean reads of grain and leaf-stem samples were mixed and assembled as reference sequences. Then the two libraries were mapped individually using RSEM software. The total numbers of mapped reads were 722255 in the grain and 1091068 in the leaf-stem. To measure gene expression, the read count was converted into RPKM, a normalized measure of read density that allows transcript levels to be compared both within and between samples. There were 48869 annotated genes yielded after removing overlapping sequences, ranging from 100 to $\geqslant$2000 bp, which provided abundant data for the analysis of grain and leaf-stem. Expression of these genes is summarized in Fig. 1.

To obtain statistical confirmation of differential gene expression, the above 48869 unigenes were chosen for further analysis. The read count was used as an indicator to identify the DEGs for the samples with no biological replications. Putative DEGs were identified using criteria mentioned in the methods. A total of 22045 DEGs were identified in both libraries, including 7355 upregulated and 14690 downregulated unigenes in the grain (Fig. 2, Appendix B, Table S2).

3.4 Identification and annotation of potential DEGs

To better understand the significance of DEGs in wheat grain, 7355 upregulated genes in the grain were focused on (Appendix C, Table S3). Among the 7355 DEGs, 6899 were assigned to 3727 GO terms (Appendix D, Table S4), providing an overview of the ontology content. Among them, 60 GO terms, the top 20 of three main categories (Fig. 3), were significantly represented using a corrected *P*-value $\leqslant$0.05 as the threshold. The terms dominant in the biological process category were carbohydrate metabolic process (GO: 0005975) and cellular component organization or biogenesis (GO: 0071840). Intracellular organelle (GO: 0043229) was dominant in the cellular component category. Alpha-amylase inhibitor activity (GO: 0015066) and hydrolase activity acting on glycosyl bonds (GO: 0016789) were dominant in the molecular function category.

KEGG pathway enrichment analysis was performed to better elucidate the functions of the upregulated DEGs. The KEGG enrichment results were displayed in a graphical form. The degree of KEGG pathway enrichment was represented by the enrichment factor, the *Q*-value and the number of unigenes enriched in a KEGG pathway. The enrichment factor indicates the ratio of DEGs enriched in this pathway to the total number of annotated unigenes in this pathway. DEGs showing upregulated expression were assigned to 212 metabolic pathways (Appendix E, Table S5) and the top 20 were chosen for further analysis. For upregulated DEGs, biosynthesis of secondary metabolites (500 DEGs), protein processing in endoplasmic reticulum (223 DEGs), ribosome (185 DEGs), and starch and sucrose metabolism (161 DEGs) were the dominant pathways (Fig. 4).

3.5 Unigenes related to protein process

In the GO significant enrichment analysis of the upregu-

Table 1 Statistical summary of cDNA sequences generated using the Roche/454

Item	Grain	Leaf-stem	Total
Number of raw reads	1067337	1519674	–
Number of clean reads	1061274	1516564	–
Number of unigenes	–	–	61393
Average length of unigenes/bp	–	–	1456
N50 of unigenes/bp	–	–	1814
N90 of unigenes/bp	–	–	750
Min length of unigenes/bp	–	–	102
Max length of unigenes/bp	–	–	12033
Total length of unigenes/bp	–	–	89401619

Table 2 Annotation statistics of unigenes

	Number of unigenes	Percentage/%
Annotated in Nr	55609	90.57
Annotated in Nt	55780	90.85
Annotated in KEGG	11876	19.34
Annotated in SwissProt	42820	69.74
Annotated in Pfam	42689	69.53
Annotated in GO	46345	75.48
Annotated in KOG	28186	45.91
Annotated in all databases	7883	12.84
Annotated in at least one database	58724	95.65
Total unigenes	61393	–

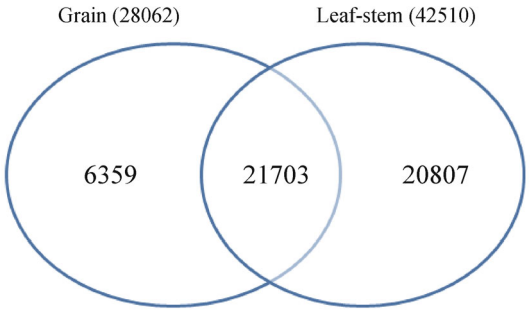

Fig. 1 Venn diagram showing the genes expressed in seed and leaf-stem tissues. The number indicated the unigenes in different tissues.

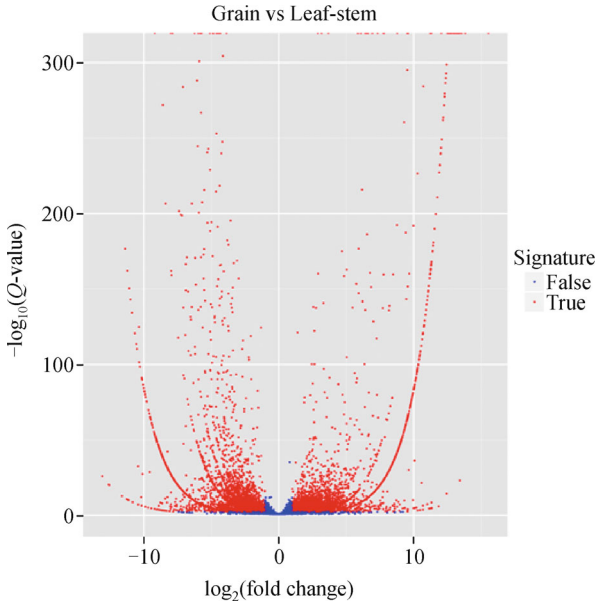

Fig. 2 Volcano diagram of DEGs. The *x*-axis indicates the fold change in unigene expression in grain and leaf-stem, and the *y*-axis indicates the statistical significance of the variance in unigene expression. A larger $-\log_{10}$ (*Q*-value) indicates that the difference is more significant. The dots in the diagram represent different unigenes. Each unigene is represented by a blue dot if the difference in expression is not significant and by a red dot if the difference is significant.

lated unigenes in grain, six GO terms which were dominant in the main molecular function category represented various enzymes associated with protein metabolism, including endopeptidase inhibitor activity, peptidase regulator activity and others. The GO term of intracellular non-membrane-bound organelle was also strongly represented. Ribosome is one of the non-membrane-bound organelles and the location of protein maturation and processing. In the KEEG pathway enrichment analysis, protein processing in the endoplasmic reticulum was strongly represented. All these showed that proteins were over-represented among the unigenes, which indicated that many proteins were synthesized and accumulated in grains.

Some unigenes were chose to verify via qPCR. Predicted high molecular weight gluten subunit (unigene: isotig52107) and gliadin (unigene: isotig08198) were found to be expressed significantly in the grain. Analysis of qPCR showed that these two kinds of protein were expressed strongly in grain at 15–20 dpa, but at a much lower level in the leaf-stem (Fig. 5a). The same result was observed for some transcription factors, *SPA* (unigene: isotig43997), wheat prolamin box binding factor (*WPBF*) (unigene: isotig37169), *GAmyb* (unigene: isotig21650) (Fig. 5b).

3.6 Unigenes associated carbohydrates

From the GO enrichment analysis of the upregulated unigenes in grain, the dominant terms in the main biological process category were almost metabolic processes associated with carbohydrates, including the sucrose metabolic process, glucan metabolic process, starch metabolic process and others. In the molecular

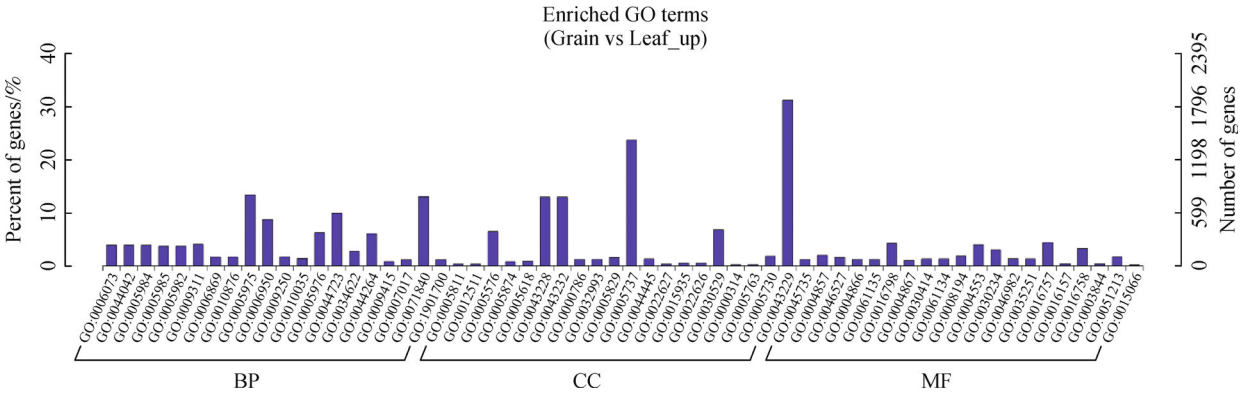

Fig. 3 GO enrichment analysis of upregulated DEGs. The results are summarized in three main categories: biologic process (BP), cellular component (CC) and molecular function (MF). The *x*-axis indicates GO terms (the description refer to Appendix D), while the *y*-axis indicates the number of genes in a category.

function category, 10 of the 20 GO terms indicated significant enrichment of multiple enzymes associated with saccharometabolism, including glucosyltransferase, sucrose synthase and 1, 4-alpha-glucan branching enzyme. The KEGG pathway enrichment analysis of genes upregulated in grain also demonstrated that starch and sucrose metabolism were significantly enriched. As above, genes associated with carbohydrate metabolism were strongly expressed.

4 Discussion

The Roche/454 GS FLX sequencer has the great strength of producing an average read length of 400 bp per sample (per bead). Compared with Illumina and SOLiD platforms, this is more efficient for *de novo* sequencing for plants without a reference genome. Although the physical map of *Triticum urartu* and *Aegilops tauschii* [27,28], the wheat D- and A-genome progenitor, has been accomplished, current knowledge of bread wheat genome sequence is insufficient for functional genomics analysis. Because of the lack of a reference sequence for wheat, the clean reads were *de novo* assembled in the study.

The metabolic pathways of the starch, protein and lipid components are clear, but the mechanisms by which they are regulated remain poorly understood. In this study, a transcript profile was obtained and many key genes associated with grain qualities can be identified to study the genetic regulatory mechanisms governing metabolism of nutritionally important substance. Consistent with the function of the endosperm as a storage organ for protein and carbohydrate, these genes encoded mainly storage proteins, enzymes involved in starch metabolism and unknown proteins.

During the wheat grain filling period, an abundance of storage proteins is synthesized and accumulated in the proteinoplast, which illustrates that an efficient regulation network exists in the wheat endosperm. There are three main types of grain seed endosperm-specific promoter conservative cis-motifs: GCN4, PB box and 5′-AACA/TA-3′, which interact with transcription factors of the alkaline leucine zipper, single zinc finger protein and R2R3MYB binding domain types [29].

Two transcription factors were selected for further study, *SPA* and *WPBF*. *SPA* is a key regulator of transcription of wheat grain storage protein, gluten strength and grain hardness [30]. *SPA* is located in the first homologous group of wheat and it belongs to the *Opaque2* transcription factor subfamily [31–33]. *WPBF* belongs to the DNA binding with one finger transcription factor subfamily. The gene encoding WPBF is located in the fifth homologous group of wheat and its product has a single zinc finger domain. WPBF regulates the expression of endosperm functional genes via binding to the prolamin box of endosperm storage protein [34], which affects the synthesis and accumulation of wheat prolamin.

Genes associated with carbohydrate metabolism were strongly expressed, which was consistent with the fact that starch content in wheat grain accounts for about 53%–70% of the total mass. Intriguingly, in the GO enrichment analysis, one GO term annotated as alpha-amylase inhibitor activity in the molecular function category was the most strongly enriched, with >1000 RPKM. Alpha-amylase inhibitor is a glucoside hydrolase, which is rich in seeds of cereal crops and leguminous crops [35]. In the endosperm these proteins might serve primarily as defense

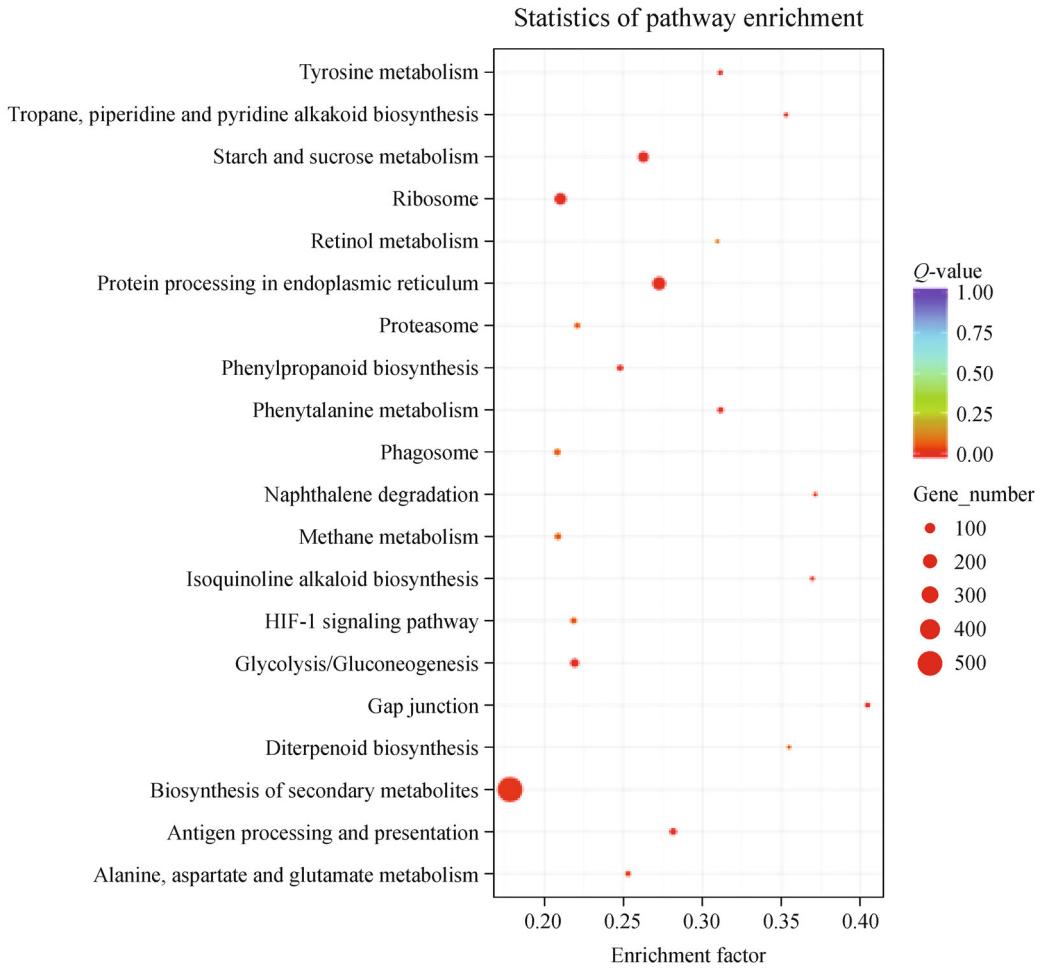

Fig. 4 KEGG pathway enrichment scatter diagram of upregulated DEGs. Only the top 20 most strongly represented pathways were displayed in the diagram. The degree of KEGG pathway enrichment was represented by an enrichment factor, the *Q*-value, and the number of unigenes enriched in a KEGG pathway. The enrichment factor indicates the ratio of differential expression unigenes enriched in this pathway to the total number of annotated unigenes in this pathway. The names of the KEGG pathways are listed along the *y*-axis. The *Q*-value indicates the corrected *P*-value, ranging from 0 and 1, and a *Q*-value closer to 0 indicates more enrichment.

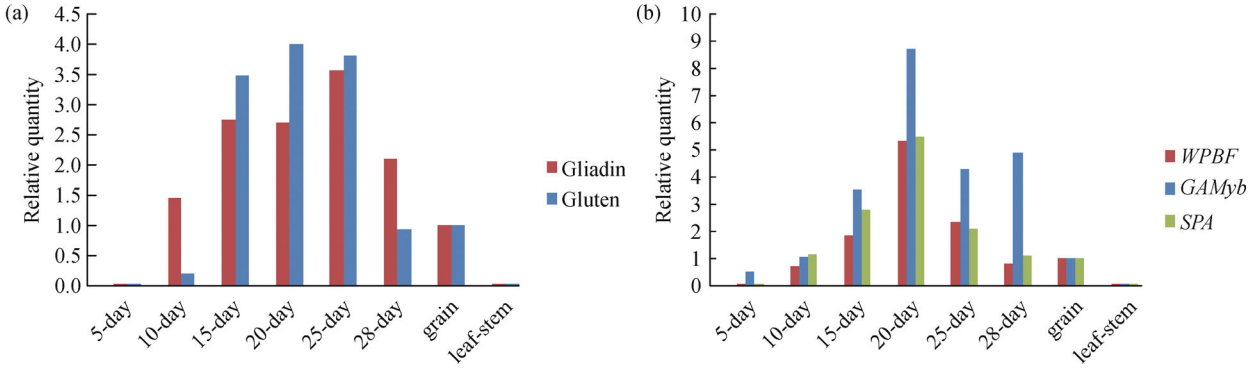

Fig. 5 Expression analysis of some upregulated genes via qPCR. (a) Genes associated with wheat storage protein, gliadin and gluten; (b) transcription factors regulating storage protein, *WPBF*, *GAMyb*, *SPA*.

proteins to protect the seeds from predatory insects [36,37]. The genes encoding these proteins have been applied in genetic engineering to improve insect resistance [38].

A total of 2669 unigenes had no significant matches to any known protein, which might be due in part to their

novelty, high divergence or untranslated regions. Moreover, there were many putative proteins of unknown function among the annotated unigenes.

This study provides the first direct performance comparison between two major different organs in bread wheat. Flag leaf, stem below the spike and the developing grain constitute the relationship between source, sink and flow. Transcriptome profiling of these at the grain filling stage in wheat will be helpful in understanding the genes involved the resource flow from flag leaves to the grains.

Overall, transcriptome analysis of gene expression profiles for grain during grain filling in wheat will permit elucidation of the genetic regulation of this process. Therefore, further research to better understand the molecular mechanisms that regulate these metabolic pathways will be warranted.

5 Conclusions

In this study, global characterization of the transcriptome of grain and leaf-stem tissues in wheat was achieved using Roche/454 GS-FLX sequencing. There were about 1.06 and 1.52 G clean-sequencing reads from grain and leaf-stem generated respectively. The enrichment of GO terms and KEGG pathway analysis were used to describe the overall biological processes upregulated in the grain. The transcriptional patterns of genes involved in protein and carbohydrate metabolism were described as examples of DEG families. These data will facilitate gene discovery and functional genomics studies in wheat.

Acknowledgements This work was supported by two grants from the National Nature Science Foundation of China (31371607 and 31071412), and a grant from Hi-Tech Research and Development Program of China (2012AA101105).

Supplementary material The online version of this article at http://dx.doi.org/(doi:10.15302/J-FASE-2014024) contains supplementary material (Appendix A – E).

Compliance with ethics guidelines Liu Wei, Zhihui Wu, Yufeng Zhang, Dandan Guo, Yuzhou Xu, Weixia Chen, Haiying Zhou, Mingshan You and Baoyun Li declare that they have no conflict of interest or financial conflicts to disclose.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

1. Gupta P K, Mir R R, Mohan A, Kumar J. Wheat genomics: present status and future prospects. *International Journal of Plant Genomics*, 2008, 2008: 1–36
2. Evers T, Millar S. Cereal grain structure and development: some implications for quality. *Journal of Cereal Science*, 2002, 36(3): 261–284
3. Laudencia-Chingcuanco D L, Stamova B S, Lazo G R, Cui X, Anderson O D. Analysis of the wheat endosperm transcriptome. *Journal of Applied Genetics*, 2006, 47(4): 287–302
4. Hammond-Kosack M C U, Holdsworth M J, Bevan M W. *In vivo* footprinting of a low molecular weight glutenin gene (LMWG-1D1) in wheat endosperm. The *EMBO Journal*, 1993, 12(2): 545–554
5. Shewry P R, Halford N G. Cereal seed storage proteins: structures, properties and role in grain utilization. *Journal of Experimental Botany*, 2002, 53(370): 947–958
6. Ball S G, Morell M K. From bacterial glycogen to starch: understanding the biogenesis of the plant starch granule. *Annual Review of Plant Biology*, 2003, 54(1): 207–233
7. Hattori J, Ouellet T, Tinker N A. Wheat EST sequence assembly facilitates comparison of gene contents among plant species and discovery of novel genes. *Genome*, 2005, 48(2): 197–206
8. Leader D J. Transcriptional analysis and functional genomics in wheat. *Journal of Cereal Science*, 2005, 41(2): 149–163
9. McIntosh S, Watson L, Bundock P, Crawford A, White J, Cordeiro G, Barbary D, Rooke L, Henry R. SAGE of the developing wheat caryopsis. *Plant Biotechnology Journal*, 2007, 5(1): 69–83
10. Wan Y, Poole R L, Huttly A K, Toscano-Underwood C, Feeney K, Welham S, Gooding M J, Mills C, Edwards K J, Shewry P R, Mitchell R A. Transcriptome analysis of grain development in hexaploid wheat. *BMC Genomics*, 2008, 9(1): 121–136
11. Yang X, Xu H, Li W, Li L, Sun J, Li Y, Yan Y, Hu Y. Screening and identification of seed-specific genes using digital differential display tools combined with microarray data from common wheat. *BMC Genomics*, 2011, 12(1): 513–524
12. Xu H, Gao Y, Wang J. Transcriptomic analysis of rice (*Oryza sativa*) developing embryos using the RNA-Seq technique. *PLoS ONE*, 2012, 7(2): e30646
13. Margulies M, Egholm M, Altman W E, Attiya S, Bader J S, Bemben L A, Berka J, Braverman M S, Chen Y J, Chen Z, Dewell S B, Du L, Fierro J M, Gomes X V, Godwin B C, He W, Helgesen S, Ho C H, Irzyk G P, Jando S C, Alenquer M L, Jarvie T P, Jirage K B, Kim J B, Knight J R, Lanza J R, Leamon J H, Lefkowitz S M, Lei M, Li J, Lohman K L, Lu H, Makhijani V B, McDade K E, McKenna M P, Myers E W, Nickerson E, Nobile J R, Plant R, Puc B P, Ronan M T, Roth G T, Sarkis G J, Simons J F, Simpson J W, Srinivasan M, Tartaro K R, Tomasz A, Vogt K A, Volkmer G A, Wang S H, Wang Y, Weiner M P, Yu P, Begley R F, Rothberg J M. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 2005, 437(7057): 376–380
14. Mardis E R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 2008, 24(3): 133–141
15. Vera J C, Wheat C W, Fescemyer H W, Frilander M J, Crawford D L, Hanski I, Marden J H. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, 2008, 17(7): 1636–1647
16. Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z, Miller W, Lipman D J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 1997, 25(17): 3389–3402
17. Finn R D, Mistry J, Tate J, Coggill P, Heger A, Pollington J E, Gavin O L, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer E L, Eddy S R, Bateman A. The Pfam protein families database. *Nucleic Acids Research*, 2010, 38(Suppl 1): D211–D222

18. Götz S, García-Gómez J M, Terol J, Williams T D, Nagaraj S H, Nueda M J, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 2008, 36(10): 3420–3435
19. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 2008, 36(Database issue): D480–D484
20. Li B, Dewey C N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 2011, 12(1): 323–338
21. Mortazavi A, Williams B A, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 2008, 5(7): 621–628
22. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology*, 2010, 11(10): R106
23. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 2010, 26(1): 136–138
24. Young M D, Wakefield M J, Smyth G K, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 2010, 11(2): R14
25. Mao X, Cai T, Olyarchuk J G, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, 2005, 21(19): 3787–3793
26. Livak K J, Schmittgen T D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ Method. *Methods*, 2001, 25(4): 402–408
27. Ling H Q, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, Gao C, Wu H, Li Y, Cui Y, Guo X, Zheng S, Wang B, Yu K, Liang Q, Yang W, Lou X, Chen J, Feng M, Jian J, Zhang X, Luo G, Jiang Y, Liu J, Wang Z, Sha Y, Zhang B, Wu H, Tang D, Shen Q, Xue P, Zou S, Wang X, Liu X, Wang F, Yang Y, An X, Dong Z, Zhang K, Zhang X, Luo M C, Dvorak J, Tong Y, Wang J, Yang H, Li Z, Wang D, Zhang A, Wang J. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature*, 2013, 496 (7443): 87–90
28. Luo M C, Gu Y Q, You F M, Deal K R, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, Jorgensen C M, Zhang Y, McGuire P E, Pasternak S, Stein J C, Ware D, Kramer M, McCombie W R, Kianian S F, Martis M M, Mayer K F, Sehgal S K, Li W, Gill B S, Bevan M W, Simková H, Dolezel J, Weining S, Lazo G R, Anderson O D, Dvorak J. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *Proceedings of the National Academy of Sciences of the United States of America*, 2013, 110(19): 7940–7945
29. Kawakatsu T, Yamamoto M P, Touno S M, Yasuda H, Takaiwa F. Compensation and interaction between RISBZ1 and RPBF during grain filling in rice. *The Plant Journal*, 2009, 59(6): 908–920
30. Ravel C, Martre P, Romeuf I, Dardevet M, El-Malki R, Bordes J, Duchateau N, Brunel D, Balfourier F, Charmet G. Nucleotide polymorphism in the wheat transcriptional activator *Spa* influences its pattern of expression and has pleiotropic effects on grain protein composition, dough viscoelasticity, and grain hardness. *Plant Physiology*, 2009, 151(4): 2133–2144
31. Albani D, Hammond-Kosack M C, Smith C, Conlan S, Colot V, Holdsworth M, Bevan M W. The wheat transcriptional activator SPA: a seed-specific bZIP protein that recognizes the GCN_4-like motif in the bifactorial endosperm box of prolamin genes. *Plant Cell*, 1997, 9(2): 171–184
32. Conlan R S, Hammond-Kosack M, Bevan M. Transcription activation mediated by the bZIP factor SPA on the endosperm box is modulated by ESBF-1 *in vitro*. *The Plant Journal*, 1999, 19(2): 173–181
33. Guillaumie S, Charmet G, Linossier L, Torney V, Robert N, Ravel C. Colocation between a gene encoding the bZip factor SPA and an eQTL for a high-molecular-weight glutenin subunit in wheat (*Triticum aestivum*). *Genome*, 2004, 47(4): 705–713
34. Dong G, Ni Z, Yao Y, Nie X, Sun Q. Wheat Dof transcription factor WPBF interacts with *TaQM* and activates transcription of an alpha-gliadin gene during wheat seed development. *Plant Molecular Biology*, 2007, 63(1): 73–84
35. Gibbs B F, Alli I. Characterization of a purified alpha-amylase inhibitor from white kidney bean (*Phaseolus vulgaris*). *Food Research International*, 1998, 31(3): 217–225
36. Feng G H, Richardson M, Chen M S, Kramer K J, Morgan T D, Reeck G R. α-amylase inhibitors from wheat: amino acid sequences and patterns of inhibition of insect and human α-amylases. *Insect Biochemistry and Molecular Biology*, 1996, 26(5): 419–426
37. Franco O L, Rigden D J, Melo F R, Bloch C Jr, Silva C P, Grossi de Sá M F. Activity of wheat α-amylase inhibitors towards bruchid α-amylases and structural explanation of observed specificities. *European Journal of Biochemistry*, 2000, 267(8): 2166–2173
38. Lajolo F M, Finardi F F. Partial characterization of the amylase inhibitor of black beans (*Phaseolus vulgaris*), variety Rico 23. *Journal of Agricultural and Food Chemistry*, 1985, 33(1): 132–138