



OPEN

DATA DESCRIPTOR

# A multi-omic resource of wheat seed tissues for nutrient deposition and improvement for human health

Jingjing Zhi<sup>1,3</sup>, Jian Zeng<sup>2,3</sup>, Yaqiong Wang<sup>1,3</sup>, Hongyan Zhao<sup>1</sup>, Guoli Wang<sup>1</sup>, Jing Guo<sup>2</sup>, Yuesheng Wang<sup>1</sup>, Mingjie Chen<sup>1</sup>, Guangxiao Yang<sup>1</sup>, Guangyuan He<sup>1</sup>, Xiaoyuan Chen<sup>2</sup>✉, Junli Chang<sup>1</sup>✉ & Yin Li<sup>1</sup>✉

As a globally important staple crop, wheat seeds provide us with nutrients and proteins. The trend of healthy dietary has become popular recently, emphasizing the consumption of whole-grain wheat products and the dietary benefits. However, the dynamic changes in nutritional profiles of different wheat seed regions (*i.e.*, the embryo, endosperm and outer layers) during developmental stages and the molecular regulation have not been well studied. Here, we provide this multi-omic resource of wheat seeds and describe the generation, technical assessment and preliminary analyses. This resource includes a time-series RNA-seq dataset of the embryo, endosperm and outer layers of wheat seeds and their corresponding metabolomic dataset, covering the middle and late stages of seed development. Our RNA-seq experiments profile the expression of 63,708 genes, while the metabolomic data includes the abundance of 984 metabolites. We believe that this was the first reported transcriptome and metabolome dataset of wheat seeds that helps understand the molecular regulation of the deposition of beneficial nutrients and hence improvements for nutritional and processing quality traits.

## Background & Summary

Wheat is one of the “big three” cereals that dominate global staple food production. Wheat seeds are milled to remove brans, leaving most starchy endosperm to produce refined white flour, which are mainly starch and proteins. Wheat has become such a popular staple crop probably because of two major reasons. First, wheat is one of the most important sources of total calorie intake and contributes significantly to our daily plant-sourced proteins, fibres, mineral nutrients and beneficial phytochemicals (*i.e.*, bioactive compounds)<sup>1</sup>. Second, wheat seeds contain a unique set of seed storage proteins (*i.e.*, glutenins and gliadins), which form polymeric protein networks as the physico-chemical basis of unique visco-elastic properties, allowing wheat dough made into numerous types of flour foods consumed globally, such as breads.

High yield and environmental adaptability have long been the main targets of wheat breeding<sup>2</sup>. In recent years, the conceptual changes of “eat full” to “eat well” have become popular in food consumption, emphasizing the nutrition value and dietary benefits of our daily meal. Nutritional studies provide evidence that long-term intake of refined white-flour foods (*e.g.*, white flour breads) as the staple is not beneficial for human health and is associated with the prevalence and development of chronic diseases, such as cardiovascular diseases and type-2 diabetes<sup>1</sup>. By contrast, consumption of wholegrain wheat-based food products has become the new dietary trend, as wheat bran provides many health-beneficial compounds<sup>3</sup>. Thereafter, increasing efforts have been made in the basic and applied research areas to understand the genetic basis of wheat quality, especially the nutritional quality traits, and to genetically improve these traits<sup>4,5</sup>. For example, several key enzymes in the carotenoid biosynthetic pathway have been functionally studied, including  $\beta$ -hydroxylase (HYD)<sup>6</sup>, lycopene epsilon cyclase (LCY $\epsilon$ )<sup>7</sup>, carotenoid cleavage dioxygenases (CCD)<sup>8</sup> and aldehyde oxidase (AO)<sup>9,10</sup>, and transgenic wheat lines or mutant lines have been investigated to discover their

<sup>1</sup>The Genetic Engineering International Cooperation Base of Chinese Ministry of Science and Technology, The Key Laboratory of Molecular Biophysics of Chinese Ministry of Education, College of Life Science and Technology, Huazhong University of Science & Technology, Wuhan, 430074, China. <sup>2</sup>Guangdong Provincial Key Laboratory of Utilization and Conservation of Food and Medicinal Resources in Northern Region, Henry Fok School of Biology and Agriculture, Shaoguan University, Shaoguan, Guangdong, 512005, China. <sup>3</sup>These authors contributed equally: Jingjing Zhi, Jian Zeng, Yaqiong Wang. ✉e-mail: [chenxy2@163.com](mailto:chenxy2@163.com); [cjl@hust.edu.cn](mailto:cjl@hust.edu.cn); [yinli2021@hust.edu.cn](mailto:yinli2021@hust.edu.cn)

effects on biofortification of  $\beta$ -carotene in wheat<sup>11–15</sup>. More recently, owing to the technological advances in metabolomics (e.g., the widely targeted metabolomics based on ultra-performance liquid chromatography–electrospray ionization–tandem mass spectrometry (UPLC–ESI–MS/MS)<sup>16</sup>), large-scale identification and quantification of phytochemicals in wheat seeds has become possible. Thus, the metabolomes of a few wheat cultivars with black, purple or blue grain colors have been studied, revealing the metabolic basis of grain color formation and possible molecular mechanisms<sup>17–21</sup>.

Based on the above-mentioned studies, it is demonstrated that: (1) the human-beneficial phytochemicals are deposited in wheat seeds in a spatial and temporal manner; (2) pigmented wheat varieties offer high levels of bioactive compounds (flavonoids, phenolics, vanillin, and azelaic acid, for instance). Many of the bioactive compounds are accumulated in the outer layers of wheat seeds and their contents vary between wheat varieties, which not only serve as the basis supporting the benefits of whole-grain wheat consumption, but, unfortunately, are mostly removed during the grain milling process<sup>22</sup>. Our group previously showed that the contents of major carotenoids in wheat (lutein, zeaxanthin and  $\beta$ -carotene) were drastically decreased during seed development, possibly representing an intrinsic limitation in carotenoid biofortification of wheat seeds<sup>14</sup>.

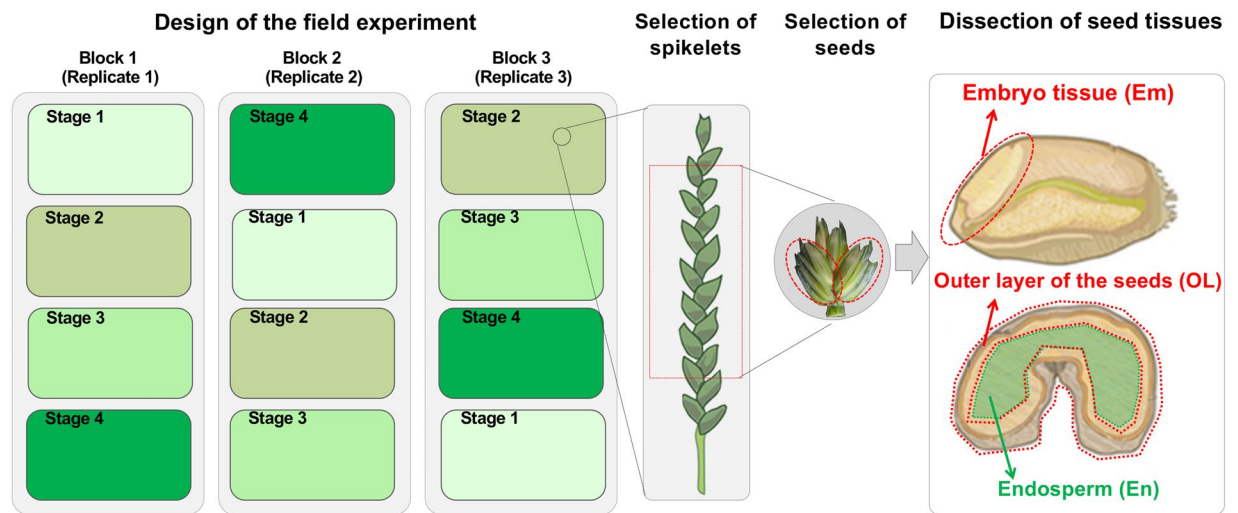
Similar to the phytochemical deposition in seeds, the seed storage proteins (SSPs) of wheat are accumulated in spatial-temporal patterns as well. These SSPs include a large portion of gluten proteins, such as high-molecular-weight glutenin subunit (HMW-GS), low-molecular-weight glutenin subunit (LMW-GS), gliadins, puroindolines and avenin-like proteins (ALPs)<sup>23</sup>. The expression of these SSP-encoding genes are dramatically up-regulated from 12 to 20 days after pollination (DAP) and peak at the middle and late stages of endosperm development (20–30 DAP)<sup>24–26</sup>. In addition, these SSPs are accumulated in different seed regions: HMW-GS and  $\gamma$ -gliadin are primarily deposited in the central endosperm cells, while the S-rich prolamins (LMW-GS,  $\alpha$ -gliadin and ALPs) are abundant in the sub-aleurone layer and the adjacent cells<sup>27–29</sup>. The spatial-temporal expression of SSP-encoding genes are tightly controlled by a combination of different transcription factors, some of which have been functionally characterized with recent advancements in wheat genomics and genetics<sup>28,30–33</sup>.

In the well-studied Poaceae species (rice and maize), complex networks of transcription factors and coregulators are known to be involved in the regulation of storage proteins and accumulation of starch and other phytochemicals, serving as, at least partly, the genetic basis of yield and nutritional and processing quality traits<sup>34,35</sup>. Compared to rice and maize, the post-genomic era of wheat studies has come until the very recent burst of Triticeae genomic resources, landmarking by the release of high-quality reference genomes of common wheat, durum wheat and their diploid ancestors<sup>36–41</sup>. While the recent accumulation of Triticeae genomic datasets, high quality transcriptomic and metabolomic resources of wheat seeds with focus on the spatial-temporal regulation of proteins and phytochemicals are scarce. The expression atlas of wheat includes a few samples of aleurone and endosperm<sup>42</sup>, while the other public RNA-seq datasets of wheat seeds emphasize on embryogenesis<sup>43</sup>. Also, environmental factors are known to contribute to grain development and metabolite abundance<sup>9,44</sup>, suggesting the necessity of a well-designed field experiment in generating transcriptome and metabolome datasets simultaneously for studying seed development. In the present work, we describe a high-quality, 310-GB RNA-seq data of wheat seed tissues together with a metabolomic dataset produced from the same samples. This multi-omic resource covers the stages and tissues suitable for studying the nutrient deposition and protein accumulation during late grain development. We believe that this resource will be helpful for understanding the spatial-temporal patterns of seed proteins and beneficial phytochemicals, for unraveling the regulatory networks involved in these metabolic processes and for wheat quality improvement.

## Methods

**Plant materials and field experiments.** The bread wheat (*Triticum aestivum* L.) cultivar L88-31 was used in the field experiment, which was conducted at the experimental field of Huazhong University of Science and Technology in Wuhan, China during the 2020–2021 season using a randomized completely block design with three replicates (Fig. 1). L88-31 belongs to a set of wheat near isogenic lines that have been significantly contributed to our understanding on wheat dough property and widely used for producing transgenic wheat lines<sup>45–47</sup>. Each block was used to collect the samples of four stages for a biological replicate and the sampling stage was randomly assigned to a plot within each block. Each plot consisted of twelve rows which were 1.5 m long and had 35 seeds per row. Since the focus of our research is to capture the dynamics of gene expression and metabolites during the late stages of wheat seed development, the four stages were used: 20, 25, 30 and 35 days after pollination (DAP). Within each plot, the main tillers and the corresponding spikes were tagged and used for sample collection. The middle part of each sampled spikes was used for seed collection and dissection of tissue samples to avoid samples from different flowering time and seed developmental status. All the spikes were collected in the morning (between 9:00 A.M. to 11:00 A.M.) to avoid potential influences of circadian on transcriptome and metabolome. Once collected from the field, the spikes were transferred to the lab and dissected on ice with scalpels and tweezers immediately to embryos, the out layer of seeds and endosperm tissues (abbreviated as Em, OL, and En, respectively, and used hereafter), followed by snap frozen in liquid nitrogen. Around 150–200 seeds from the same plot were sampled and pooled together to form a biological replicate (including the tissues for both RNA-seq and metabolomics) (Fig. 1). Owing to the tissue availability, the embryo tissues collected at 20, 25, 30, and 35 DAP (Em20, Em25, Em30, and Em35, respectively) were used for RNA-seq analysis and other experiments in future, while the endosperm and outer layer tissues collected at 20, 25, 30, and 35 DAP (En20, En25, En30, En35, and OL20, OL25, OL30, OL35, respectively) were used for both metabolomics and RNA-seq, with the remaining tissues saved for future experiments.

**RNA extraction, library construction and sequencing.** Total RNA was extracted with TRIzol reagent. The quality of extracted RNA samples was examined by agarose gel electrophoresis, NanoDrop 2000,



**Fig. 1** Overview of the experimental design and sample collection for RNA-seq and metabolomics.

Sample Name	RNA Conc.	volume	quantity	A260/280	A260/230	RIN
En20-1	342	35	11.97	1.92	1.64	8.7
En20-2	404	35	14.14	1.88	1.51	8.4
En20-3	444	35	15.54	1.92	1.75	8.6
En25-1	356	35	12.46	1.87	1.32	8.7
En25-2	165	35	5.78	1.93	1.68	8.5
En25-3	252	35	8.82	1.94	1.56	8.6
En30-1	228	35	7.98	1.93	1.75	8.8
En30-2	304	35	10.64	1.92	1.67	8.2
En30-3	196	35	6.86	1.94	1.71	8.4
En35-1	118	35	4.13	1.94	1.72	8.6
En35-2	230	35	8.05	1.96	1.74	8.7
En35-3	218	35	7.63	1.93	1.68	8.3
OL20-1	478	35	16.73	1.91	1.72	8.8
OL20-2	421	35	14.74	1.94	1.74	8.9
OL20-3	387	35	13.55	1.93	1.76	8.7
OL25-1	365	35	12.78	1.95	1.68	8.9
OL25-2	358	35	12.53	1.92	1.62	8.5
OL25-3	401	35	14.04	1.96	1.63	8.6
OL30-1	369	35	12.92	1.92	1.71	8.4
OL30-2	374	35	13.09	1.93	1.68	8.7
OL30-3	362	35	12.67	1.94	1.74	8.6
OL35-1	241	35	8.44	1.92	1.62	8.8
OL35-2	231	35	8.09	1.96	1.68	8.7
OL35-3	242	35	8.47	1.91	1.72	8.4
Em20-1	231	35	8.09	1.91	1.72	8.1
Em20-2	189	35	6.62	1.94	1.74	8.0
Em20-3	187	35	6.55	1.93	1.76	7.7
Em25-1	265	35	9.28	1.95	1.68	7.9
Em25-2	158	35	5.53	1.92	1.62	7.5
Em25-3	201	35	7.04	1.96	1.63	7.6
Em30-1	269	35	9.42	1.92	1.71	8.3
Em30-2	174	35	6.09	1.93	1.68	7.7
Em30-3	162	35	5.67	1.94	1.74	7.6
Em35-1	241	35	8.44	1.92	1.62	7.8
Em35-2	231	35	8.09	1.96	1.68	8.7
Em35-3	182	35	6.37	1.91	1.72	8.4

**Table 1.** The quality of each RNA sample. RIN, RNA integrity.

and Agilent 2100 Bio-analyzer (Table 1). Standard protocols for the Illumina NovaSeq platform were used for construction of the wheat mRNA libraries. RNA-seq libraries were sequenced to generate 150-bp pair-end reads. For sequence quality control, cutadapt (<https://cutadapt.readthedocs.io/en/stable/>)<sup>48</sup> and FASTX-Toolkit ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)) were used to trim low-quality base pairs from the 3' end of each sequence and the quality of raw and clean data was checked with FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). The quality-filtered clean reads were mapped to the wheat reference genome of Chinese Spring (IWGSC\_v1.0) using HISAT2 v2.0.1-beta with default parameters (<https://daehwankimlab.github.io/hisat2/>)<sup>39,49</sup>. Only uniquely-mapped reads were retained and the read counts that aligned to the 110,790 gene models annotated of the wheat reference genome were calculated by using the featureCount software (<https://subread.sourceforge.net/featureCounts.html>)<sup>50</sup>. FPKM (fragments per kilobase of exon per million mapped sequence reads) values were calculated for each gene model. Genes met the following criteria were considered as expressed in a stage: (1) at least 5 reads mapped to a gene in each of the three replicates; (2) the average FPKM at a stage should be  $\geq 0.5$ . The Pearson correlation coefficients between biological replicates were calculated using gene expression values. The differentially expressed genes (DEGs) were determined with edgeR by comparing the read count data between stages within each tissue or by comparing the tissues of each stages (<https://bioconductor.org/packages/release/bioc/html/edgeR.html>)<sup>51,52</sup>. The significance threshold for edgeR-based differentially expressed genes was fold change  $\geq 2$  and a false discovery rate (FDR)-adjusted  $P$ -value  $< 0.05$ .

**Metabolomic analysis.** To capture the metabolomic changes during wheat seed development, quasi-targeted metabolomics was employed to analyze the wheat samples under a contract with Novogene Co., Ltd. (Beijing, China). The metabolomic method was established by Novogene and reported previously<sup>53–55</sup>. The method is described in detail as below.

*Metabolites extraction.* Wheat tissues (100 mg per sample) were grounded individually into powder with liquid nitrogen and the homogenate was well vortexed together with prechilled 500  $\mu$ L 80% methanol. The samples were then subject to a 5-min cold incubation (on ice) followed by centrifugation (15,000 g, 4 °C for 20 min). The supernatant was diluted with LC-MS grade water to a final concentration of 53% methanol. The samples were subsequently transferred to a new Eppendorf tube followed by centrifugation (15,000 g, 4 °C for 20 min). After that, the supernatant was injected into the LC-MS/MS system<sup>56,57</sup>.

*HPLC-MS/MS analysis.* LC-MS/MS analyses were conducted with an ExionLC™ AD system (SCIEX) coupled with a QTRAP® 6500+ mass spectrometer (SCIEX). By using a 20-min linear gradient at a flow rate of 0.4 mL/min for the positive/negative polarity mode, the sample was injected onto a Xselect HSS T3 (2.1  $\times$  150 mm, 2.5  $\mu$ m) with 0.1% formic acid-water and 0.1% formic acid-acetonitrile as the eluent A and B, respectively<sup>54</sup>. The following parameters for solvent gradient was used: 2% B, 2 min; 2–100% B, 15.0 min; 100% B, 17.0 min; 100–2% B, 17.1 min; 2% B, 20 min. QTRAP® 6500+ mass spectrometer was operated in the positive polarity mode with curtain gas of 35 psi, collision gas of medium, ionspray voltage of 5500 V, temperature of 550 °C, ion source gas of 1:60, and ionsource gas of 2:60. QTRAP® 6500+ mass spectrometer was operated in the negative polarity mode with the following settings: curtain gas of 35 psi, collision gas of medium, ionspray voltage of –4500 V, temperature of 550 °C, ion source gas of 1:60, ion source gas of 2:60.

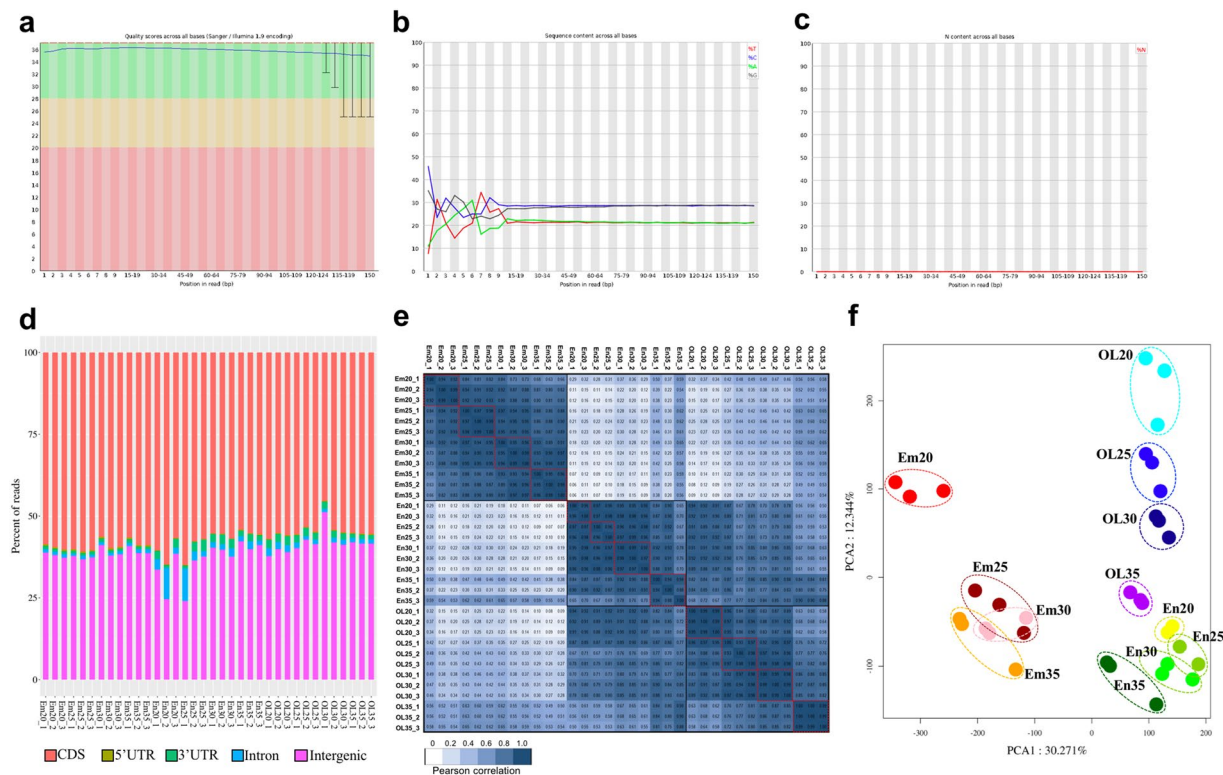
*Metabolites identification and quantification.* Metabolite identification was based on the in-house database using MRM (Multiple Reaction Monitoring)<sup>58</sup>. In the in-house library of Novogene Co., Ltd., more than 2500 commercially available purified standard compounds have been registered to the LC-MS/MS platform for determination of their characteristics. This in-house database includes 190+ amino acids and derivatives, 200+ organic acids and derivatives, 100+ nucleotide and derivatives, more than 120 flavonoid compounds, more than 40 anthocyanins, over 30 vitamin compounds, 30+ alkaloids and derivatives, 20+ phenolamides, 95+ carbohydrates and derivatives, 100+ lipids (including phospholipids, fatty acyls, glycerophospholipids and glycerolipids), 20+ phytohormone compounds, 600+ known compounds from medicinal plants and over 1000 other compounds (such as anthracenes, benzene and substituted derivatives, cinnamic acids and derivatives, coumarins and derivatives, iridoid derivatives, etc.).

To accurately identify biochemicals, retention time (RT) with a narrow RT window, accurate mass match to the library entries ( $\pm 0.005$  amu), Q1 (parent ion) and Q3 and the MS/MS forward and reverse scores between the experimental data and the authentic standards were applied as the criteria<sup>59,60</sup>. SCIEX OS (version 1.4) was used to process the HPLC-MS/MS data files to integrate and correct the peak with the following parameters: minimum peak height, 500; signal/noise ratio, 5; gaussian smooth width, 1. The area of each peak represents the relative content of the corresponding substance.

*Metabolomic data analysis.* These identified metabolites were annotated using Human Metabolome Database (HMDB) database<sup>61</sup>, Lipidmaps database<sup>62</sup> and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database<sup>63</sup>. Principal components analysis (PCA) and partial least squares discriminant analysis (PLS-DA) were conducted with the metabolomics software metaX<sup>64</sup>. We applied univariate analysis ( $t$ -test) to calculate the statistical significance ( $P$ -value). The following criteria was used to identify differential metabolites: variable important in projection value VIP  $> 1$  and  $P$ -value  $< 0.05$ ,  $\log_2$ (fold change)  $\geq 1$  or  $\leq -1$ .

## Data Records

This data set contains two parts: the first part is the RNA-seq data and the derived gene expression data; the second part is the results of metabolite identification and quantification with metabolomics. Both parts use the same set of sample names, in which En, Em and OL stands for the endosperm, embryo, and outer layer tissues, respectively, with two digits standing for seed developmental stages and the suffix standing for biological



**Fig. 2** Quality control and clustering analysis of the RNA-seq data set. **(a)** Quality score of per position in the reads. **(b)** Sequence content per base of the clean data. **(c)** N content per base of the clean data. **(d)** The percent of reads mapped to coding sequences (CDS), 5'untranslated regions (5'UTR), 3'untranslated regions (3'UTR), introns and intergenic regions for each RNA-seq sample. **(e)** Hierarchical clustering analysis. **(f)** Primary component analysis (PCA) of the RNA-seq samples. The analyzed results of Em20\_1 sample are shown as a representative in the (a–c).

replicates. For the RNA-seq data, the raw sequencing data containing all 36 samples were deposited in the NCBI (Bioproject ID: PRJNA891918; SRA experiments No. SRP430408)<sup>65</sup>. The gene expression matrix and the results of differential expression analysis are available at figshare<sup>66</sup>. For the metabolomic data, the metabolite information (*i.e.*, compound name, formula, exact Q1, molecular weight, RT and CAS No.), relative quantities and differential metabolite results have been made available by Novogene Co., Ltd. and are deposited at figshare<sup>67,68</sup>. Owing to the contract with Novogene, the raw metabolomic data have not been publicly available. Other data related to this data set (such as RNA quality, RNA-seq statistics) are provided in the manuscript (Table 1, Supplementary Tables S1–S3).

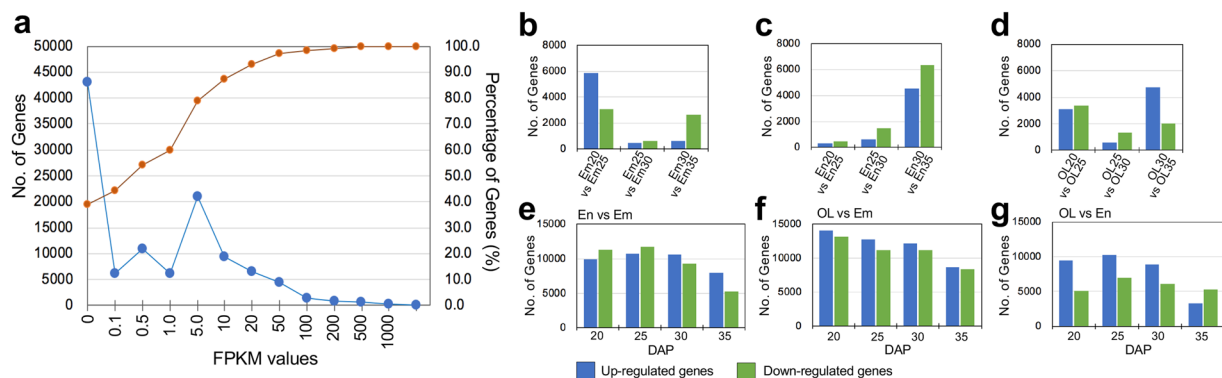
## Technical Validation

**Quality control.** The quality of extracted total RNA is key to high-quality RNA-seq libraries and successful downstream experiments, because low RNA integrity (RIN) values may affect the quality of RNA-seq libraries and may lead to potential deviations in gene expression. In our project, we ensured that the RIN values of all RNA samples were > 7.0. The quality parameters of each RNA samples are shown in Table 1.

## Quality Validation

**Quality validation for the RNA-seq samples.** To ensure that the transcriptome and metabolome data obtained meet statistical standards, we have designed the experiment in consideration of several potential influencing factors, including replicates of the fields, growth differences between major tillers and smaller, minor tillers, differences in the developmental rate between the seeds from upper, middle and lower parts of the spike, and growth differences between individual plants and seeds (Fig. 1). In our experiments, we sampled several hundreds of seeds for each sample and the seeds were dissected on ice into three tissues: embryo (Em), the out layers (OL) and endosperm (En).

The RNA-seq experiments obtained high sequencing quality scores and high clean read ratios. The Q30 scores range from 87.67% to 95.49% and the clean read ratios ranges from 97.18% to 99.37% (Supplementary Table S1; Fig. 2a–c; samples' metadata in Supplementary Table S2). The GC content of all RNA-seq samples was relatively stable, varying from 48.95% to 57.89% (Supplementary Table S1). The percentage of RNA-seq reads mapped to different parts of coding regions demonstrate that a large portion (from 45.4% to 60.84%) of the reads were mapped to coding sequence (CDS) with a small fraction of the reads mapped to 5' and 3' untranslated regions (0.12% to 0.54% for 5'UTR and 0.77% to 2.94% for 3'UTR, respectively) for most of the RNA-seq

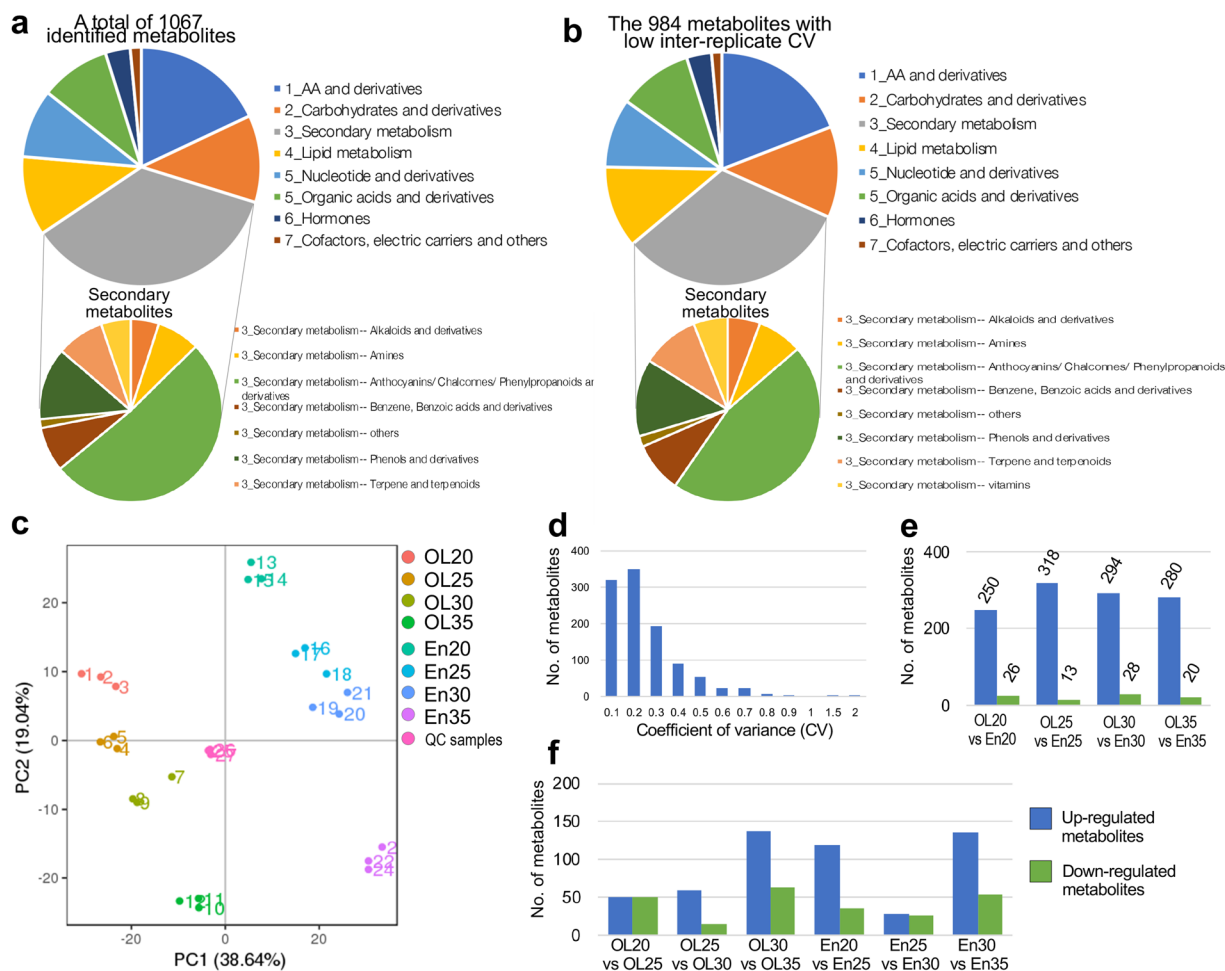


**Fig. 3** Gene expression of the RNA-seq samples (a) and differential expression analyses between the stages (b–d) or the tissue types (e–g). (a) The distribution of gene expression values (in FPKM) using Em20\_1 as a representative. (b–d) The number of DEGs identified between the stages from embryo (b), endosperm (c) and the outer layers (d). (e–g) The number of DEGs identified between the tissue types from the same stages.

samples (Fig. 2d). However, two samples, En20-2 and En25-1, might be skewed compared to the other samples, as they contained 9.51% and 9.92% percentage of reads, respectively, mapped to intronic regions, indicating DNA contamination (Fig. 2d). Consistent with this result, En20-2 and En25-1 RNA samples indeed had lower A260/230 scores (1.51 and 1.32, respectively; Table 1) and lower GC contents (49.45% and 48.95%, respectively; Supplementary Table S1). The lowered A260/230 and GC content of En20-2 and En25-1 samples also indicate DNA contamination in these samples. In addition, sample OL30-1 had a higher percentage of reads mapped to intergenic regions (Fig. 2d). Taken together with its low mapping rate but relatively acceptable uniquely mapped rate and the global expression pattern correlated well with the corresponding replicates, we think OL30-1 can be acceptable for gene expression analysis (Fig. 2e,f; Supplementary Table S1).

Differential expression analysis and the expression values were calculated for the high-confidence geneIDs annotated in the wheat reference genome (Supplementary Table S1, Table S2). Among the 110,790 annotated genes, a large portion had expression values ranging from 1 to 100 in the FPKM unit for each RNA-seq sample (a representative FPKM distribution per sample shown in Fig. 3a), supporting FPKM > 0.5 as a reasonable threshold for the expressed genes. The number of expressed genes varied from 37,862 to 51,368 (Supplementary Table S4), consistent with previous reports of wheat seeds samples<sup>43</sup>. The results of correlation heatmap and principle component analysis (PCA) based on the expression between the replicates further support that En20-2 and En25-1 samples are deviated from the remaining RNA-seq samples (Supplementary Figure S1). Thus, En20-2 and En25-1 samples are excluded from the identification of differentially expressed genes (DEGs). The number of DEGs between each pair of RNA-seq samples are shown (Fig. 3, Supplementary Table S4). Also, when En20-2 and En25-1 were excluded, the remaining RNA-seq samples could be well separated in the correlation heatmap and PCA plot (Fig. 2e,f, respectively). In the PCA result, the principle components 1 (PC1) and 2 (PC2) explains 30.27% and 12.34% of the variations, respectively, and well corresponds to the tissue types and developmental stages. The PCA results show: (1) the three tissue types can be clearly separated; (2) the outer layer and embryo tissues exhibit greater changes in gene expression during the four stages compared to the endosperm tissues. Indeed, the number of DEGs between the three tissues support this finding, with at least over 5000 up- or down-regulated genes identified at each of the four stage (Fig. 3e–g). By contrast, the number of DEGs between different stages within a particular seed tissue are much less, ranging from several hundreds to a few thousands. For example, only a few hundred genes were up- or down-regulated in the comparison of En20-En25 and En25-En30, consistent with the relative stable transcriptomic status of the endosperm samples as indicated in the PCA results (Figs. 2f, 3). In brief, the results of PCA and differential expression analysis consistently reflect the transcriptomic differences between the seed tissues and stages, demonstrating that this RNA-seq dataset is of high quality to study the gene expression regulation in different seed tissues during developmental stages.

**Quality validation for the metabolomic samples.** The metabolomic analysis identified 1,067 known metabolites in the samples of wheat endosperm (En) and seed outer layers (OL), covering a diverse set of compounds, including amino acids, carbohydrates, lipids, nucleotides, organic acids and their derivatives and many metabolites belong to secondary metabolic pathways (Fig. 4a). By calculating the coefficient of variance of each metabolite within the three replicates, 984 metabolites<sup>67,68</sup> with low CV values (<0.8) have been kept (CV distribution shown in Fig. 4d and Supplementary Fig. 2, meta-data for the metabolomic samples in Supplementary Table S2), with their metabolic categories given in Fig. 4b. PCA analysis well detected the metabolic differences among the samples, with PC1 and PC2 explaining 38.64% and 19.04% metabolic variations, respectively, and probably corresponding to the differences between the tissues and the stages (Fig. 4c). Partial least squares discriminant analysis (PLS-DA) was performed to identify the metabolites with differential abundance between the samples (Fig. 4e,f, Supplementary Table S5).



**Fig. 4** Characterization of the metabolomic data. **(a)** Classification of the 1,067 identified metabolites. **(b)** Classification of the 984 metabolites with low coefficient of variance values within the replicates. **(c)** PCA analysis of the metabolic samples. **(d)** The distribution of coefficient of variance values within the replicates for each metabolite (the three replicates of OL20 were used as a representative). **(e)** The number of differential metabolites between outer layer samples and endosperm samples at each of the four stages. **(f)** The number of differential metabolites between the stages within each the outer-layer or endosperm tissue.

Interestingly, around 250 to 300 metabolites were found to be up-regulated in the outer layer samples at each of the four stages (Fig. 4e), consistent with previous results that the outer layers of wheat seeds have more metabolites with higher abundance<sup>22</sup>.

### Code availability

No custom code was generated for this work.

Received: 2 November 2022; Accepted: 3 April 2023;

Published online: 10 May 2023

### References

- Hazard, B. *et al.* Strategies to improve wheat for human health. *Nat Food* **1**, 475–480, <https://doi.org/10.1038/s43016-020-0134-6> (2020).
- Kiszonas, A. M. & Morris, C. F. Wheat breeding for quality: a historical review. *Cereal Chem* **95**, 17–34, <https://doi.org/10.1094/CCHEM-05-17-0103-FI> (2017).
- McRae, M. P. Health benefits of dietary whole grains: an umbrella review of meta analyses. *J Chiropractic Med* **16**, 10–18, <https://doi.org/10.1016/j.jcm.2016.08.008> (2017).
- Zhai, S., Xia, X. & He, Z. Carotenoids in staple cereals: metabolism, regulation, and genetic manipulation. *Front Plant Sci* **7**, 1197, <https://doi.org/10.3389/fpls.2016.01197> (2016).
- Yu, X., Luo, Q., Huang, K., Yang, G. & He, G. Prospecting for microelement function and biosafety assessment of transgenic cereal plants. *Front Plant Sci* **9**, 326, <https://doi.org/10.3389/fpls.2018.00326> (2018).
- Qin, X., Zhang, W., Dubcovsky, J. & Tian, L. Cloning and comparative analysis of carotenoid  $\beta$ -hydroxylase genes provides new insights into carotenoid metabolism in tetraploid (*Triticum turgidum* ssp. *durum*) and hexaploid (*Triticum aestivum*) wheat grains. *Plant Mol Biol* **80**, 631–646, <https://doi.org/10.1007/s11103-012-9972-4> (2012).
- Rodriguez-Suarez, C., Mellado-Ortega, E., Hornero-Mendez, D. & Atienza, S. G. Increase in transcript accumulation of *Psy1* and *e-Lcy* genes in grain development is associated with differences in seed carotenoid content between durum wheat and tritordeum. *Plant Mol Biol* **84**, 659–673, <https://doi.org/10.1007/s11103-013-0160-y> (2014).

8. Qin, X., Fischer, K., Yu, S., Dubcovsky, J. & Tian, L. Distinct expression and function of carotenoid metabolic genes and homoeologs in developing wheat grains. *BMC Plant Biol* **16**, 155, <https://doi.org/10.1186/s12870-016-0848-7> (2016).
9. Colasuonno, P. *et al.* Characterization of *Aldehyde Oxidase* (AO) genes involved in the accumulation of carotenoid pigments in wheat grain. *Front Plant Sci* **8**, 863, <https://doi.org/10.3389/fpls.2017.00863> (2017).
10. Yu, S. & Tian, L. Assessing the role of carotenoid cleavage dioxygenase 4 homoeologs in carotenoid accumulation and plant growth in tetraploid wheat. *Front Nutr* **8**, 740286, <https://doi.org/10.3389/fnut.2021.740286> (2021).
11. Cong, L. *et al.* Expression of phytoene synthase1 and carotene desaturase *CrtI* genes result in an increase in the total carotenoids content in transgenic elite wheat (*Triticum aestivum* L.). *J Agric Food Chem* **57**, 8652–8660, <https://doi.org/10.1021/jf9012218> (2009).
12. Wang, C. *et al.* Enrichment of provitamin A content in wheat (*Triticum aestivum* L.) by introduction of the bacterial carotenoid biosynthetic genes *CrtB* and *CrtI*. *J Exp Botany* **65**, 2545–2556, <https://doi.org/10.1093/jxb/eru138> (2014).
13. Zeng, J. *et al.* Metabolic engineering of wheat provitamin A by simultaneously overexpressing *CrtB* and silencing carotenoid hydroxylase (*TaHYD*). *J Agric Food Chem* **63**, 9083–9092, <https://doi.org/10.1021/acs.jafc.5b04279> (2015).
14. Zeng, J. *et al.* The lycopene  $\beta$ -cyclase plays a significant role in provitamin A biosynthesis in wheat endosperm. *BMC Plant Biol* **15**, 112, <https://doi.org/10.1186/s12870-015-0514-5> (2015).
15. Yu, S., Li, M., Dubcovsky, J. & Tian, L. Mutant combinations of lycopene  $\epsilon$ -cyclase and  $\beta$ -carotene hydroxylase 2 homoeologs increased  $\beta$ -carotene accumulation in endosperm of tetraploid wheat (*Triticum turgidum* L.) grains. *Plant Biotechnol J* **20**, 564–576, <https://doi.org/10.1111/pbi.13738> (2022).
16. Chen, W. *et al.* A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: Application in the study of rice metabolomics. *Mol Plant* **6**, 1769–1780, <https://doi.org/10.1093/mp/sst080> (2013).
17. Zhai, S. *et al.* A genome-wide association study reveals a rich genetic architecture of flour color-related traits in bread wheat. *Front Plant Sci* **9**, 1136, <https://doi.org/10.3389/fpls.2018.01136> (2018).
18. Wang, X. *et al.* Metabolomics and gene expression analysis reveal the accumulation patterns of phenylpropanoids and flavonoids in different colored-grain wheats (*Triticum aestivum* L.). *Food Chem* **138**, 109711, <https://doi.org/10.1016/j.foodres.2020.109711> (2020).
19. Shi, T. *et al.* Metabolomics analysis and metabolite-agronomic trait associations using kernels of wheat (*Triticum aestivum*) recombinant inbred lines. *Plant J* **103**, 279–292, <https://doi.org/10.1111/tj.14727> (2020).
20. Li, L. *et al.* Transcriptomics integrated with widely targeted metabolomics reveals the mechanism underlying grain color formation in wheat at the grain-filling stage. *Front Plant Sci* **12**, 757750, <https://doi.org/10.3389/fpls.2021.757750> (2021).
21. Wang, F. *et al.* Metabolomics and transcriptomics provide insights into anthocyanin biosynthesis in the developing grains of purple wheat (*Triticum aestivum* L.). *J Agric Food Chem* **63**, 11171–11184, <https://doi.org/10.1021/acs.jafc.1c01719> (2021).
22. Zhu, A. *et al.* Metabolomic analysis of the grain pearling fractions of six bread wheat varieties. *Food Chem* **369**, 130881, <https://doi.org/10.1016/j.foodchem.2021.130881> (2022).
23. Wang, D., Li, F., Cao, S. & Zhang, K. Genomic and functional genomics analyses of gluten proteins and prospect for simultaneous improvement of end-use and health-related traits in wheat. *Theor Appl Genet* **133**, 1521–1539, <https://doi.org/10.1007/s00122-020-03557-5> (2020).
24. Moore, K. *et al.* The dynamics of protein body formation in developing wheat grain. *Plant Biotechnol J* **14**, 1876–1882, <https://doi.org/10.1111/pbi.12549> (2016).
25. Gao, Y. *et al.* The endosperm-specific transcription factor TaNAC019 regulates glutenin and starch accumulation and its elite allele improves wheat grain quality. *Plant Cell* **33**, 603–622, <https://doi.org/10.1093/plcell/koaa040> (2021).
26. Shewry, P. R. & Lafandra, D. Wheat glutenin polymers I. structure, assembly and properties. *J Cereal Sci* **106**, 103486, <https://doi.org/10.1016/j.jcs.2022.103486> (2022).
27. Tosi, P. *et al.* Trafficking of storage proteins in developing grain of wheat. *J Exp Botany* **60**, 979–991, <https://doi.org/10.1093/jxb/ern346> (2009).
28. Tosi, P., Gritsch, C. S., He, J. & Shewry, P. R. Distribution of gluten proteins in bread wheat (*Triticum aestivum*) grain. *Ann Bot* **108**, 23–35, <https://doi.org/10.1093/aob/mcr098> (2011).
29. He, J. *et al.* Spatial patterns of gluten protein and polymer distribution in wheat grain. *J Agric Food Chem* **61**, 6207–6215, <https://doi.org/10.1021/jf401623d> (2013).
30. Luo, G. *et al.* The MYB family transcription factor TuODORANT1 from *Triticum urartu* and the homolog TaODORANT1 from *Triticum aestivum* inhibit seed storage protein synthesis in wheat. *Plant Biotechnol J* **19**, 1863–1877, <https://doi.org/10.1111/pbi.13604> (2021).
31. Shen, L. *et al.* A novel NAC family transcription factor SPR suppresses seed storage protein synthesis in wheat. *Plant Biotechnol J* **19**, 992–1007, <https://doi.org/10.1111/pbi.13524> (2021).
32. Luo, G. *et al.* Genome-wide identification of seed storage protein gene regulators in wheat through coexpression analysis. *Plant J* **108**, 1704–1720, <https://doi.org/10.1111/tj.15538> (2021).
33. Li, J. *et al.* TaNAC100 acts as an integrator of seed protein and starch synthesis exerting pleiotropic effects on agronomic traits in wheat. *Plant J* **108**, 829–840, <https://doi.org/10.1111/tj.15485> (2021).
34. Li, Q. & Wu, Y. The encyclopedia of maize kernel gene expression. *J Integr Plant Biol* **62**, 879–881, <https://doi.org/10.1111/jipb.12869> (2020).
35. Dai, D., Ma, Z. & Song, R. Maize endosperm development. *J Integr Plant Biol* **63**, 613–627, <https://doi.org/10.1111/jipb.13069> (2021).
36. Zhao, G. *et al.* The *Aegilops tauschii* genome reveals multiple impacts of transposons. *Nat Plants* **3**, 946–955, <https://doi.org/10.1038/s41477-017-0067-8> (2017).
37. Luo, M. C. *et al.* Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **551**, 498–502, <https://doi.org/10.1038/nature24486> (2017).
38. Ling, H. *et al.* Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature* **557**, 424–428, <https://doi.org/10.1038/s41586-018-0108-0> (2018).
39. IWGSC. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191, <https://doi.org/10.1126/science.aar7191> (2018).
40. Maccaferri, M. *et al.* Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* **51**, 885–895, <https://doi.org/10.1038/s41588-019-0381-3> (2019).
41. Walkowiak, S. *et al.* Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283, <https://doi.org/10.1038/s41586-020-2961-x> (2020).
42. Ramírez-González, R. H. *et al.* The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089, <https://doi.org/10.1126/science.aar6089> (2018).
43. Xiang, D. *et al.* The transcriptional landscape of polyploid wheats and their diploid ancestors during embryogenesis and grain development. *Plant Cell* **31**, 2888–2911, <https://doi.org/10.1105/tpc.19.00397> (2019).
44. Hill, C. B. *et al.* Whole-genome mapping of agronomic and metabolic traits to identify novel quantitative trait loci in bread wheat grown in a water-limited environment. *Plant Physiol* **162**, 1266–1281, <https://doi.org/10.1104/pp.113.217851> (2013).
45. Lawrence, G. J. *et al.* Dough and baking quality of wheat lines deficient in glutenin subunits controlled by the *Glu-A1*, *Glu-B1* and *Glu-D1* loci. *J Cereal Sci* **7**, 109–112, [https://doi.org/10.1016/S0733-5210\(88\)80012-2](https://doi.org/10.1016/S0733-5210(88)80012-2) (1988).



46. Barro, F. *et al.* Transformation of wheat with high molecular weight glutenin subunit genes results in improved functional properties. *Nat Biotechnol* **15**, 1295–1299, <https://doi.org/10.1038/nbt1197-1295> (1997).
47. Li, M. *et al.* Effect of extra cysteine residue of new mutant 1Ax1 subunit on the functional properties of common wheat. *Sci Rep* **7**, 7510, <https://doi.org/10.1038/s41598-017-07541-w> (2017).
48. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17**, 10–12, <https://doi.org/10.14806/ej.17.1.200> (2011).
49. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360, <https://doi.org/10.1038/nmeth.3317> (2015).
50. Yang, L., Gordon, K. S. & Wei, S. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **7**, 923–930, <https://doi.org/10.1093/bioinformatics/btt656> (2014).
51. Ali, M., Brian, A. W., Kenneth, M., Lorian, S. & Barbara, W. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628, <https://doi.org/10.1038/nmeth.1226> (2008).
52. Mark, D. R., Davis, J. M. & Gordon, K. S. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140, <https://doi.org/10.1093/bioinformatics/btp616> (2010).
53. Zhang, J. *et al.* Ovary abortion induced by combined waterlogging and shading stress at the flowering stage involves amino acids and flavonoid metabolism in maize. *Front Plant Sci* **12**, 778717, <https://doi.org/10.3389/fpls.2021.778717> (2021).
54. Peng, Z. *et al.* Integration of metabolome and transcriptome studies reveals flavonoids, abscisic acid, and nitric oxide comodulating the freezing tolerance in *Liriope spicata*. *Front Plant Sci* **12**, 764625, <https://doi.org/10.3389/fpls.2021.764625> (2022).
55. Zhang, Z. *et al.* Honeybee gut lactobacillus modulates host learning and memory behaviors via regulating tryptophan metabolism. *Nat Commun* **13**, 2037, <https://doi.org/10.1038/s41467-022-29760-0> (2022).
56. Want, E. J. *et al.* Global metabolic profiling of animal and human tissues via UPLC-MS. *Nat Protocols* **8**, 17–32, <https://doi.org/10.1038/nprot.2012.135> (2012).
57. Sellick, C. A. *et al.* Metabolite extraction from suspension cultured mammalian cells for global metabolite profiling. *Nat Protocols* **6**, 1241–1249, <https://doi.org/10.1038/nprot.2011.366> (2011).
58. Ping, L. *et al.* MRM-Ion Pair Finder: a systematic approach to transform non-targeted mode to pseudo-targeted mode for metabolomics study based on liquid chromatography-mass spectrometry. *Anal Chem* **87**, 5050–5055, <https://doi.org/10.1021/acs.analchem.5b00615> (2015).
59. Yuan, M., Breikopf, S. B., Yang, X. & Asara, J. M. A positive/negative ion-switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. *Nat Protocols* **7**, 872–881, <https://doi.org/10.1038/nprot.2012.024> (2012).
60. Barri, T. & Dragsted, L. O. UPLC-ESI-QTOF/MS and multivariate data analysis for blood plasma and serum metabolomics: effect of experimental artefacts and anticoagulant. *Anal Chim Acta* **768**, 118–128, <https://doi.org/10.1016/j.aca.2013.01.015> (2013).
61. Wishart, D. S. *et al.* HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res* **46**, D608–D617, <https://doi.org/10.1093/nar/gkx1089> (2018).
62. Hartler, J. LIPID MAPS: Tools and Databases. In: Wenk, M. (eds) *Encyclopedia of Lipidomics*. Springer, Dordrecht. [https://doi.org/10.1007/978-94-007-7864-1\\_11-1](https://doi.org/10.1007/978-94-007-7864-1_11-1) (2015).
63. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457–D462, <https://doi.org/10.1093/nar/gkv1070> (2016).
64. Wen, B., Mei, Z., Zeng, C. & Liu, S. MetaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinform* **18**, 183, <https://doi.org/10.1186/s12859-017-1579-y> (2017).
65. *NCBI Sequence Read Archive* <https://identifiers.org/ncbi/insdc.sra:SRP403408> (2023).
66. Li, Y. RNA-seq expression data for “A multi-omic resource of wheat seed tissues for nutrient deposition and improvement for human health”. *figshare*. <https://doi.org/10.6084/m9.figshare.22492360.v1> (2023).
67. Li, Y. The metabolomic data set for “A multi-omic resource of wheat seed tissues for nutrient deposition and improvement for human health”. *figshare*. <https://doi.org/10.6084/m9.figshare.22268290.v1> (2023).
68. Li, Y. Sample metadata for the study “A multi-omic resource of wheat seed tissues for nutrient deposition and improvement for human health”. *figshare*. <https://doi.org/10.6084/m9.figshare.22269151.v1> (2023).

## Acknowledgements

This work was supported by National Major Projects of China (2016ZX08010004-004), the National Natural Science Foundation of China (32272126, 31771418, 31371614, 31901537), the Natural Science Foundation of Hubei China (2016CFB549), the Fundamental Research Funds for Central Universities, HUST (2021XXJS070, 3004170157), Wuhan Knowledge Innovation Project (2022020801010073) and the Opening fund of Guangdong Provincial Key Laboratory of Utilization and Conservation of Food and Medicinal Resources in Northern Region.

## Author contributions

J.Z. (Jian Zeng), Y.W. (Yaqiong Wang), G.H., X.C., J.C. and Y.L. conceived and designed the research project. J.Z. (Jingjing Zhi), J.Z. (Jian Zeng), Y.W. (Yaqiong Wang), G.W., H.Z. and J.G. performed the experiments. Y.W. (Yuesheng Wang), M.C., G.Y. and G.H. contributed to experiment supervision and project administration. J.Z. (Jingjing Zhi), J.Z. (Jian Zeng), Y.W. (Yaqiong Wang) and Y.L. analyzed the data. J.Z. (Jingjing Zhi), J.Z. (Jian Zeng), Y.W. (Yaqiong Wang), X.C., J.C. and Y.L. drafted the manuscript and all authors revised and finalized the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-023-02133-y>.

**Correspondence** and requests for materials should be addressed to X.C., J.C. or Y.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023