Wang *et al. Applied Biological Chemistry* (2024) 67:30
https://doi.org/10.1186/s13765-024-00882-x

KSABC
The Korean Society for Applied Biological Chemistry

ARTICLE

Open Access

Classification of soybeans from different habitats based on metabolomic–transcriptomic integration

Jinghui Wang[1], Qiyou Zheng[1*], Chenxu Wang[1] and Ao Zhou[1]

Abstract

Soybeans are a significant agricultural product in China, with certain geographical locations often yielding higher quality, and thus more expensive, soybean crops. In this study, metabolomics and transcriptomics analyses were conducted on soybean samples from nine regions in Heilongjiang and Liaoning Provinces using untargeted liquid chromatography–mass spectrometry (LC–MS) and Illumina sequencing technologies. The primary objective was to devise an effective and unbiased method for determining the geographical origin of each soybean variety to mitigate potential fraudulent practices. Through multidimensional and unidimensional analyses, successful identification of differentially expressed metabolites (DEMs) and differentially expressed genes (DEGs) was achieved, yielding statistically significant outcomes. Integration of the metabolomics and transcriptomics datasets facilitated the construction of a correlation network model capable of distinguishing soybeans originating from different geographical locations, leading to the identification of significant biomarkers exemplifying noteworthy distinctions. To validate the feasibility of this method in practical applications, partial least squares discriminant analysis was employed to differentiate soybean samples from the nine regions. The results convincingly showcased the applicability and reliability of this approach in accurately pinpointing the geographical origin of soybeans. Distinguishing itself from prior research in soybean traceability, this study incorporates an integrated analysis of metabolomics and transcriptomics data, thereby unveiling biomarkers that offer a more precise differentiation of soybean traits across distinct regions, thereby bridging a critical research gap within the soybean traceability domain. This innovative dual-data integration analysis methodology is poised to enhance the accuracy of soybean traceability tools and lay a new foundation for future agricultural product identification research.

Keywords Soybean, Metabolomics, Transcriptomics, Omics integration analysis, Geographic identification

Introduction

Soybean, often hailed as the "crop of the century", is a nutrient-rich legume that serves as a key ingredient in a myriad of products, ranging from medicine to cosmetics [1]. Its high content of protein, fat, dietary fiber, and isoflavones has made it a dietary staple for both humans and livestock. China, being one of the world's largest producers and consumers of soybeans, has its production areas primarily located in the northeast regions, Shaanxi, Sichuan, and the middle and lower reaches of the Yangtze River [2]. Recently, due to intricate shifts in agricultural production and distribution economics, soybeans from specific geographic locations have seen a surge in prices. This has led to an influx of fraudulent suppliers selling substandard soybean products, infringing upon consumer rights [3]. Consequently, there is an urgent need for a method that can accurately trace the origins of soybeans and other agricultural products. Such a method

*Correspondence:
Qiyou Zheng
zqy@tccu.edu.cn
[1] College of Information Technology, Jilin Agricultural University, Changchun 130118, China

would not only safeguard the geographic indicators of these products but also ensure traceability from farm to table. This would ultimately promote fair trade practices in the bulk agricultural products market.

To uphold consumer interests and maintain order in the soybean trading market, it is imperative to enforce relevant laws and develop reliable technologies for crop origin identification. Current origin identification methods involve complex analyses of variable compounds from different geographical sources, identification of effective source confirmation features, establishment of a discrimination model, and prediction of sample classifications. These analyses may include mineral element fingerprinting [4], stable isotope identification [5], near-infrared spectroscopy [6], metabolomics [7], and electrical sensing using an electronic nose [8]. However, origin indicators can be influenced by numerous factors, making single source data-based indicators (e.g., mineral elements, fatty acids) inconsistent and less reproducible. Metabolites are often considered as signals reflecting an organism's "genetic structure and adaptability to the environment" [9]. With the advent of high-throughput technology, contemporary research strategies aimed at characterizing molecular differences among plants grown in various environments primarily focus on identifying unique molecular signatures in the plants' genome, transcriptome, proteome and/or metabolome [10]. The integration of transcriptomics and metabolomics has garnered increasing attention as it provides a more comprehensive understanding of biological systems [11, 12]. Numerous studies have explored the effects of abiotic or biotic stress on plant metabolism through comparative analyses of microarray and metabolomics data. Leveraging recent advancements in experimental platforms and technology, we have embarked on studies examining the relationship between gene expression and metabolite levels in plants to gain insights into how these networks are integrated [13, 14]. Thus, our goal is to establish a more stable and accurate method for determining the point of origin.

Metabolomics, which allows for the detection of low molecular weight organic acids, fatty acids, amino acids, sugars, and other metabolites in biological samples through high-throughput screening, data processing, and integration, is currently the preferred method for analyzing differences in crop breeding strategies, plant-microbe interactions, plant agronomic traits, and quality classification [15, 16]. This approach has proven useful in differentiating the quality and content of various products such as coffee beans [17], sea cucumber [18], tea [19], and honey [1]. LC–MS is commonly used in metabolomics analysis due to its broad applicability in analyzing metabolites that are difficult to volatilize or have poor thermal stability [20]. Metabolomics not only offers high throughput approaches but also provides high resolution and sensitivity for small molecule detection [21]. Unbiased metabolomics systematically and comprehensively analyzes soybean metabolite data based on biologically relevant single metabolites, which is far more effective in determining differential metabolites (DEMs) [22–24].

Transcriptomics provides a means to examine functional genomic elements and overall gene expression profiles related to crop growth under diverse environmental conditions [25]. The integration of metabolomics and transcriptomics data has been applied to biological stress and breeding research in recent years, providing a deeper understanding of these fields than either approach alone could offer [26, 27]. Since both gene transcription and metabolism occur simultaneously in an organism, their integrated analysis provides a powerful tool for verifying unique plant molecular characteristics to ultimately determine potential origin [28, 29].

In recent research, Gong Lijuan et al. predicted the potential distribution of soybeans in frigid regions in China using MaxEnt modeling based on climate scenarios [30]. The research focused on climate-related factors influencing soybean distribution and provided insights into potential habitat changes under various climate scenarios. Sheng Cui Dong et al. analyzed the geographical specificity of fatty acid and multi-element fingerprints of soybeans in northern China to identify the geographical origin of soybean samples [31]. The research utilized gas chromatography and mass spectrometry to classify soybean samples based on their metabolic fingerprints. Nawaz Muhammad Amjad et al. focused on the geographic distribution and germplasm conservation of Korean wild soybeans (Glycine soja) as an important genetic resource for soybean improvement [32]. It highlighted the importance of conserving wild soybean germplasms and their potential role in soybean breeding programs. Sachar Silky and Anuj Kumar provided a comprehensive survey of techniques used in computer vision for the automatic identification of plants using leaf images [33]. The study discussed various feature extraction techniques and classification methods to identify different plant species, emphasizing the importance of automated plant identification for conservation purposes. Zhang Jun et al. explored the protist community assembly and ecological roles in soybean fields in different regions of China [34]. It investigated the interactions between protists, bacteria, and fungi in the bulk soil and rhizosphere of soybean plants, highlighting the ecological importance of protists in soybean fields. Yin Leikun et al. proposed an optimized feature selection strategy for mapping individual crop types in the Sanjiang Plain, China, using Sentinel-2 time series images

[35]. The study demonstrated a significant improvement in crop mapping accuracy by integrating specific features of individual crop types. Feng Xiong et al. focused on the differences in metabolites in the rhizosphere of soybeans under varying soil potassium conditions, highlighting the effects of potassium status on root exudates and metabolites in the soil [36]. The research emphasized cultivar differences in metabolites and root exudation under different potassium conditions. Xian Yuyang et al. predicted the current and future distributions of major food crop designated geographical indications (GIs) in China under climate change using the MaxEnt model [37]. The research emphasized the importance of considering climate change scenarios in predicting the potential climate suitability of food crop GIs. Lucas Kássio R. Garcia et al. evaluated biodiversity damage indicators for soybean crops in different ecoregions in Brazil, assessing the potential impacts of soybean production on biodiversity loss [38]. The research proposed adjustments to biodiversity indicators for the life cycle assessment of soybean production in different Brazilian ecoregions. Chotekajorn Awatsaya et al. evaluated seed amino acid content in wild Japanese soybean populations to assess genetic diversity and free amino acid abundance in wild soybean seeds [39]. The study highlighted the variation in amino acid content among wild soybean accessions and its implications for soybean conservation and improvement efforts. HU Yu-qi et al. investigated the sexual compatibility between transgenic soybeans and different wild soybean populations to evaluate the potential gene flow via pollen [40]. The study assessed podding and seed sets after artificial hybridization, demonstrating the compatibility of wild soybeans with transgenic soybeans. Saleem Aamir et al. analyzed genetic diversity and selective sweeps in a European soybean germplasm collection compared to Chinese soybean collections [41]. The research identified selective sweep regions related to domestication and improvement traits, emphasizing the genetic diversity available in the European soybean collection. Azizah Firdausi Nur et al. detected metabolites in the rhizosphere of soybeans under different soil potassium conditions, highlighting the impact of potassium status on root exudation and metabolites in the soil [42]. The study demonstrated cultivar differences in metabolites in the rhizosphere of soybeans under different potassium conditions. Liu Yang et al. examined the interrelationship between latitudinal differences and metabolic changes in Tilia amurensis Rupr [43]. The research analyzed metabolite profiles of T. amurensis from different latitudes, highlighting the influence of environmental factors on metabolite differences in T. amurensis at varying latitudes. Kim Myoungsub et al. focused on transcriptional changes in a Korean soybean cultivar during its interaction with Pseudomonas syringae at the late infection phase [44]. The study identified differentially expressed genes related to plant immune response and metabolic processes during the compatible interaction between soybeans and the bacterial pathogen.

Moreover, advances in other technological fields have generated new possibilities for identifying the origin of soybeans and facilitating fair trade of agricultural products. In the realm of sustainable greenhouse cultivation, Durmanov et al. conducted a comprehensive study on the pivotal role of technological advancements and management practices in promoting agricultural sustainability and their impact on crop metabolism [45]. By precisely manipulating environmental conditions during soybean growth, such as temperature, light exposure, and water availability, significant alterations in the growth and metabolic performance of soybeans can be achieved. A sustained adaptation to technological progress is crucial for maintaining competitiveness within the food and beverage industry. Suseno and Basrowi's research underscores the critical significance of technological innovation and integration in enhancing operational efficiency and market responsiveness [46]. Furthermore, Kassymbek et al.'s work emphasizes the crucial contribution of feed processing technology to the quality and efficiency of agricultural production, providing valuable insights for analyzing soybean metabolic characteristics [47].

While existing studies have made valuable contributions in predicting soybean distribution, identifying geographical origin based on metabolic fingerprints, and analyzing genetic diversity, a gap exists in the integration of metabolomic and transcriptomic datasets, limiting their ability to offer a comprehensive understanding of regional characteristics and traceability in soybeans. In contrast, our study bridges this gap by integrating metabolomic–transcriptomic data, enabling the accurate classification of soybeans from different habitats and the identification of unique gene transcripts and metabolites for reliable origin tracing while minimizing interference from planting conditions and agronomic practices. By addressing these limitations and innovating agricultural product traceability methods, our research provides a cutting-edge approach to exploring regional traits and tracing the origin of soybeans.

In this study, we used integrated metabolomic and transcriptomic datasets to discover molecular signatures unique to soybeans from different habitats. We selected Heilongjiang and Liaoning Provinces of northeast China as the research areas, with nine main soybean producing areas of different latitudes chosen as the sample sources. We employed a combination of multi-dimensional and single-dimensional analysis to screen differential metabolomic and transcriptomic data. Integrated analyses were

conducted by screening the common pathways of genes and metabolites to select key molecules of interest. The overarching aim of the study was to identify novel gene transcripts and metabolites that can inform future mechanistic research and provide a theoretical and practical basis for the traceability of soybean origin.

Materials and methods

Sample collection and preparation

Liaoning province, characterized by a temperate monsoon climate with an average annual temperature of 9.6 °C, and Heilongjiang province, known for its cold temperate monsoon climate with an average annual temperature of 4.0 °C, were selected as the regions for soybean sample collection in this study [48, 49]. The samples were provided by the local Academy of Agricultural Sciences and research institutions. In late October 2021, soybean samples were collected from nine sites in China. Each individual soybean sample weighed 20 g, from which 5.4g of high-quality, plump seeds were meticulously selected for data analysis. Specifically, the selected soybean varieties were: Xingnong 20 (XN20) from Bei'an City, Heilongjiang; Heihe 43 (HH43) from Nenjiang City, Heilongjiang; Dongsheng 22 (DS22) from Hailun City, Heilongjiang; Suinong 52 (SN52) from Bayan County, Heilongjiang; Tiedou 67 (TD67) from Zhuanghe City, Liaoning; Liaodou 36 (LD36) from Huludao City, Liaoning; Tiefeng 31 (TF31) from Jinlandian District, Liaoning; Liaodou 15 (LD15) from Xinmin City, Liaoning; and Tiedou53 (TD53) from Linghai City, Liaoning (Fig. 1). The fresh soybean samples were promptly cleaned with purified water and 75% ethanol. After removing the residual liquid, the samples were placed into an enzyme-free tube, rapidly frozen in liquid nitrogen, and stored at − 80 °C. Each sample was analyzed in triplicate.

Metabolite extraction and LC–MS analysis

The metabolic profile was analyzed using a LC–MS system (Thermo Fisher Scientific, Waltham, Massachusetts, USA), which comprised of Dionex U3000 ultra high-performance liquid chromatography (UHPLC) tandem QE high resolution mass spectrometer. The analysis was

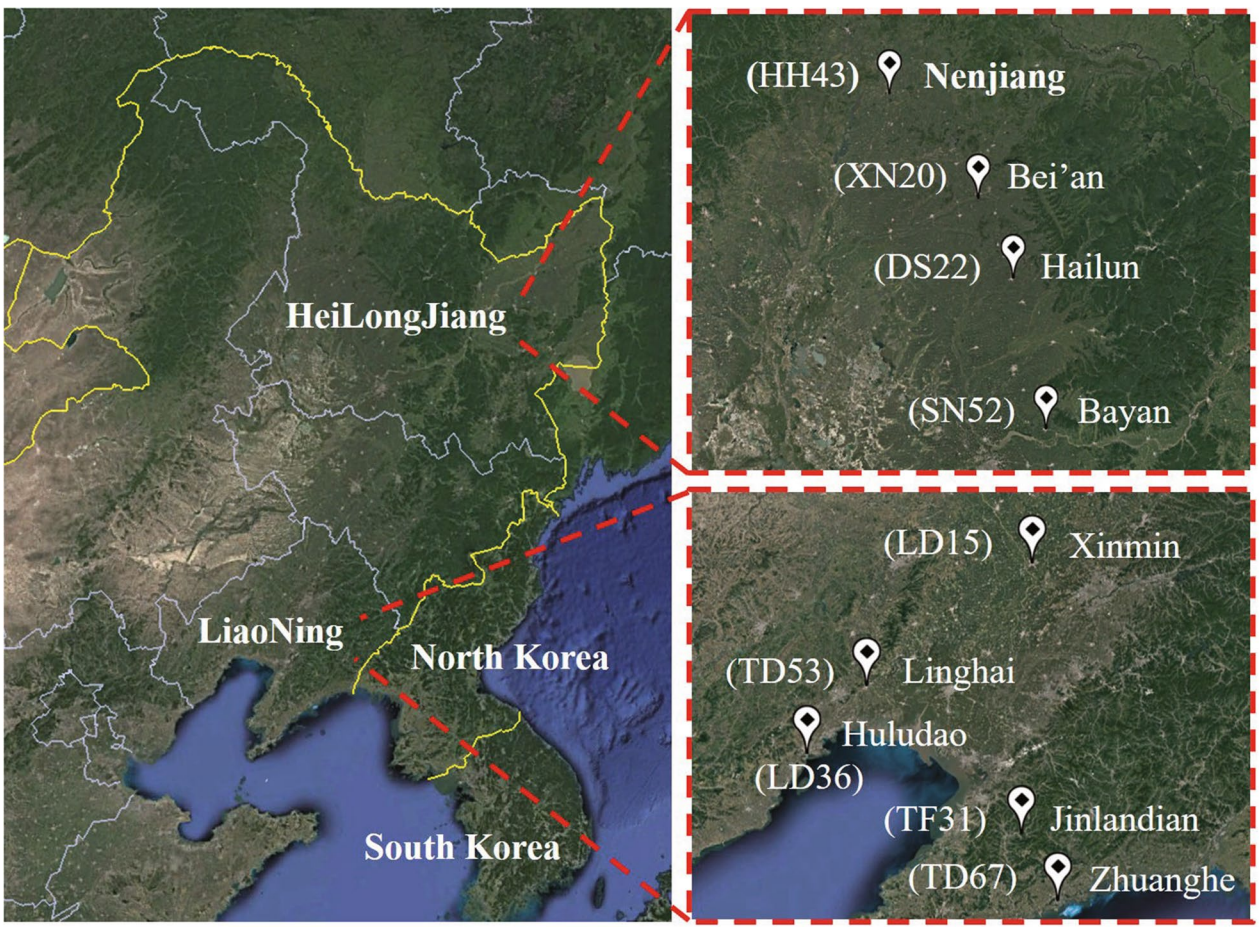

Fig. 1 Location of the production area where soybean samples were obtained

conducted under ESI positive and negative ion modes. Chromatography was performed using an ACQUITY UPLC HSS T3 (1.8 μm, 2.1 × 100 mm) column under both positive and negative modes. The binary gradient eluting solvent consisted of (A) water (containing 0.1% formic acid, vol/vol) and (B) acetonitrile (containing 0.1% formic acid, v/v), with the following components used for the separation gradient: 0 min, 5% B; 2 min, 5% B; 4 min, 25% B; 8 min, 50% B; 10 min, 80% B; 14 min, 100% B; 15 min, 100% B; 15.1 min, 5% B, and 16 min, 5% B. The flow rate was set at 0.35 mL/min with the column temperature maintained at 40 °C. All samples were kept at a temperature of 4 °C during the analysis and the final injection volume was set to be at 5 μL. The mass detected ranged from m/z of 100 to 1000. The resolution of full MS scanning was set at a value of 70,000, and the resolution of HCD-MS/MS scanning was set at 17500. The collision energy was set to values of 10, 20, and 40eV respectively. The mass spectrometer was operated as follows: spray voltage was set to be at 3800V for positive mode (+) and 3000V for negative mode (−); sheath gas flow was set to be at 35 arbitrary units (AU); auxiliary gas flow was set to be at 8 AU; capillary temperature was set to be at 320 °C while Aux gas heater temperature was set to be at 350 °C; S Lens input frequency class was set to be at 50. All the analyses were repeated three times for reproducibility.

Progenisis QI v2.3 software (Nonlinear Dynamics, Newcastle, UK) was used for baseline filtering, peak identification, integration, retention time correction, peak alignment and normalization of original LC–MS data. Main parameters assessed were: precursor tolerance, 5 ppm/10 ppm (in-house database); product tolerance: 10 ppm/20 ppm (in-house database); product ion threshold: 5%. For the extracted data, ion peaks with missing values (0) > 50% within the group were deleted, and the 0 value was replaced with half of the minimum value. Qualitatively obtained compounds were screened according to the qualitative screening score. The screening standard was 36 points (full score: 60), and qualitative results < 36 were considered inaccurate and deleted. Finally, the positive and negative ion data were combined into a data matrix table in which all the extracted data was analyzed.

RNA extraction and data preprocessing

Total RNA was extracted from each sample, and residual DNA was digested with DNase. Eukaryotic mRNA was enriched using oligo (dT) magnetic beads. The mRNA was fragmented into short segments and used as a template to synthesize the first cDNA strand with random hexamer primers. The first strand was then reverse transcribed to form double-stranded (ds) cDNA. The purified ds cDNA underwent end repair, A-tail extension, and sequencing adaptor connection. Fragment size was selected, and PCR amplification was performed. After the constructed library passed quality inspection using Agilent 2100 Bioanalyzer, Illumina HiSeqTM 2500 or HiSeq X 10 was used for sequencing to generate paired end reads of 125 bp or 150 bp.

Raw reads generated from high-throughput sequencing were in fastq format. Trimmomatic software [50] was employed for quality control, adaptor removal, and filtering out of low-quality bases and N-bases. Hisat2 [51] was used to map clean reads to the reference soybean genome under default parameters, and samples were evaluated through genome alignment rate. Gene expression was quantified using Cufflinks software to obtain FPKM values [52, 53]. Htseq-count software [54] was used to obtain the number of reads mapping onto genes within each sample, and data were normalized using the estimateSizeFactors function of DESeq (2012) R package [55, 56]. Fold change (FC) differences and statistical significance were calculated using the nbinomTest function, with significance threshold set to FDR < 0.1, and $p < 0.05$. A FC difference of > 1.2 or < 0.833 was used as the biological significance threshold. Gene ontology (GO) and KEGG enrichment analyses were performed to capture differentially expressed genes (DEGs) [57].

Results

Metabolome and transcriptome

Metabolome

A total of 21,893 discrete peaks were detected in each soybean sample from Liaoning and Heilongjiang provinces, including 11,809 peaks in the negative ion mode and 10,084 peaks in the positive ion mode. These metabolites include 2695 lipids and lipid-like molecules, 915 phenylpropanoic acids and polyketides, 762 organic oxygen compounds, 633 organic acids and derivatives, 608 organic heterocyclic compounds, 427 benzoic acids, 92 nucleosides, nucleotides and analogues, 37 alkaloids and derivatives, 37 lignin, neolignan and related compounds, 34 organic nitrogen compounds, 21 hydrocarbons, 14 organic sulfur compounds, 4 organic halogen compounds, 2 homogeneous non-metallic compounds, one homogeneous metal compound, one hydrocarbon derivative, one organic phosphorus compound, one organic 1,3 dipole compound, and 1998 unclassified compounds. The metabolite types and amounts for each sample are shown in Additional file 1: Table S1.

Transcriptome

A total of 186.54 G of clean sequencing reads was obtained using the Illumina sequencing platform. The effective data volume of each sample ranged from 6.36 to 7.4 G, the Q30 base distribution was between 90.39% and 94.89%, and the average GC content was 45.21%. In

mapping reads to the reference genome, the mapping rate ranged from 92.99 to 95.55%. Using known reference gene sequences and annotation files in the database, relative abundance of gene expression in each sample was identified by sequence similarity comparison. After the counts were obtained by comparison in htseq-count software, data was filtered to remove genes with zero reads. The total number of genes with detectable expression in soybeans from each region is shown in Additional file 1: Table S2.

Multidimensional statistical analysis

Principal component analysis (PCA) was initially performed on LC–MS and RNAseq datasets to determine outliers and trends, as shown in Additional file 1: Fig. S1. PCA confirmed that no outliers were present ($p < 0.05$). According to PCA based on Bray-Curtis distance, metabolites measured in soybeans from the same origin were of similar type and composition, while there were significant differences between the metabolites of soybeans from different origins. Metabolome data were then filtered for biomarkers using partial least squares discriminant analysis (PLS-DA), which has higher discriminatory capacity for biomarker discovery compared with PCA [58]. Orthogonal PLS-DA (OPLS-DA) was then used to filter the noise irrelevant to soybean classification, improve analytical ability and maximize the differences between soybean groups within the model. The Variable Importance in Projection (VIP) was obtained according to OPLS-DA modeling to measure the influence intensity and interpretative ability of the expression mode of each metabolite, and to mine the DEMs with biological significance. The screening criterion was that the VIP characteristics of the first principal component of the OPLS-DA model were> 1.0. However, this analysis alone was not sufficient to classify all variables accurately, because it only involved the natural clustering of samples, while the load map only provides the preliminary assumption of the variable distribution. Therefore, we conducted traditional single dimension statistical analysis on all samples to more accurately discern the differential characteristics of soybeans from each production area.

Screening of potential biomarkers

Screening DEMs and enrichment analysis of KEGG pathways

Both multi-dimensional and single-dimensional analyses were employed to screen differential metabolites (DEMs). Following the multi-dimensional analysis of the original metabolomic data, single-dimensional analysis was utilized to verify whether DEMs between groups were statistically significant. The characteristic fold change (FC) value was > 1.2 or < 0.833, and the characteristic p-value of Student's t-test was < 0.05. The number of DEMs in each group pair is shown in Additional file 1: Fig. S2.

Pathway enrichment analysis was subsequently conducted on DEMs to comprehend the different metabolic pathways activated in different soybean samples. Based on the KEGG database (https://www.kegg.jp/), the enriched metabolic pathways were analyzed for DEMs, with $p < 0.05$ as the threshold [59]. The top 20 differentially enriched metabolic pathways of DEMs are shown in Fig. 2. In sample DS22/SN52 from Heilongjiang province, significantly enriched pathways included tryptophan and galactose metabolism, citric acid cycle, and valine/leucine/isoleucine turnover. In sample HH43/SN52, significantly enriched pathways included isoflavone biosynthesis, tryptophan, glycerol phospholipid, and glycine/serine/threonine metabolism. In sample XN20/SN52, significantly enriched pathways included isoflavone and anthocyanin biosynthesis, galactose, tryptophan and cyanoamino acid, alanine/aspartate/glutamate and linoleic acid metabolism, and citric acid cycle. In sample LD15/TF31 from Liaoning province, the isoflavone biosynthesis pathway was the most highly enriched of all significant pathways, while flavone/flavonoid/flavonol biosynthesis, arginine biosynthesis, and tryptophan metabolism were also pathways that were significantly enriched. In sample LD36/TF31, highly enriched pathways included isoflavone and arginine biosynthesis, tryptophan and glycerol phospholipid metabolism. Other significantly enriched pathways included flavone/flavonoid/flavonol biosynthesis, histidine and glycine/serine/threonine metabolism, aminoacyl tRNA biosynthesis, ABC transporters, alanine/aspartic acid/glutamate metabolism, and anthocyanin biosynthesis. In sample TD53/TF31, the highly enriched pathways included isoflavone and flavonoid biosynthesis while other significantly enriched pathways were arginine biosynthesis and cyanoamino acid metabolism. In sample TD67/TF31, the highly enriched pathways included isoflavone biosynthesis, tryptophan, glycerol phospholipid and linoleic acid metabolism while other significantly enriched pathways included phenylpropanoid and arginine biosynthesis as well as autophagy-other. The top 20 enriched pathways of DEMs obtained by pairwise comparison improved the reliability of biomarker discovery and provided preliminary hypotheses for the distribution of all variables.

Comparing DEGs and KEGG pathway annotation

The soybean transcriptome from each sample was analyzed to determine differentially expressed genes (DEGs) among soybeans from different regions. The R language DESeq2 package was used to standardize the number of gene counts within each sample, in which the BaseMean value was used to estimate the expression level. The fold

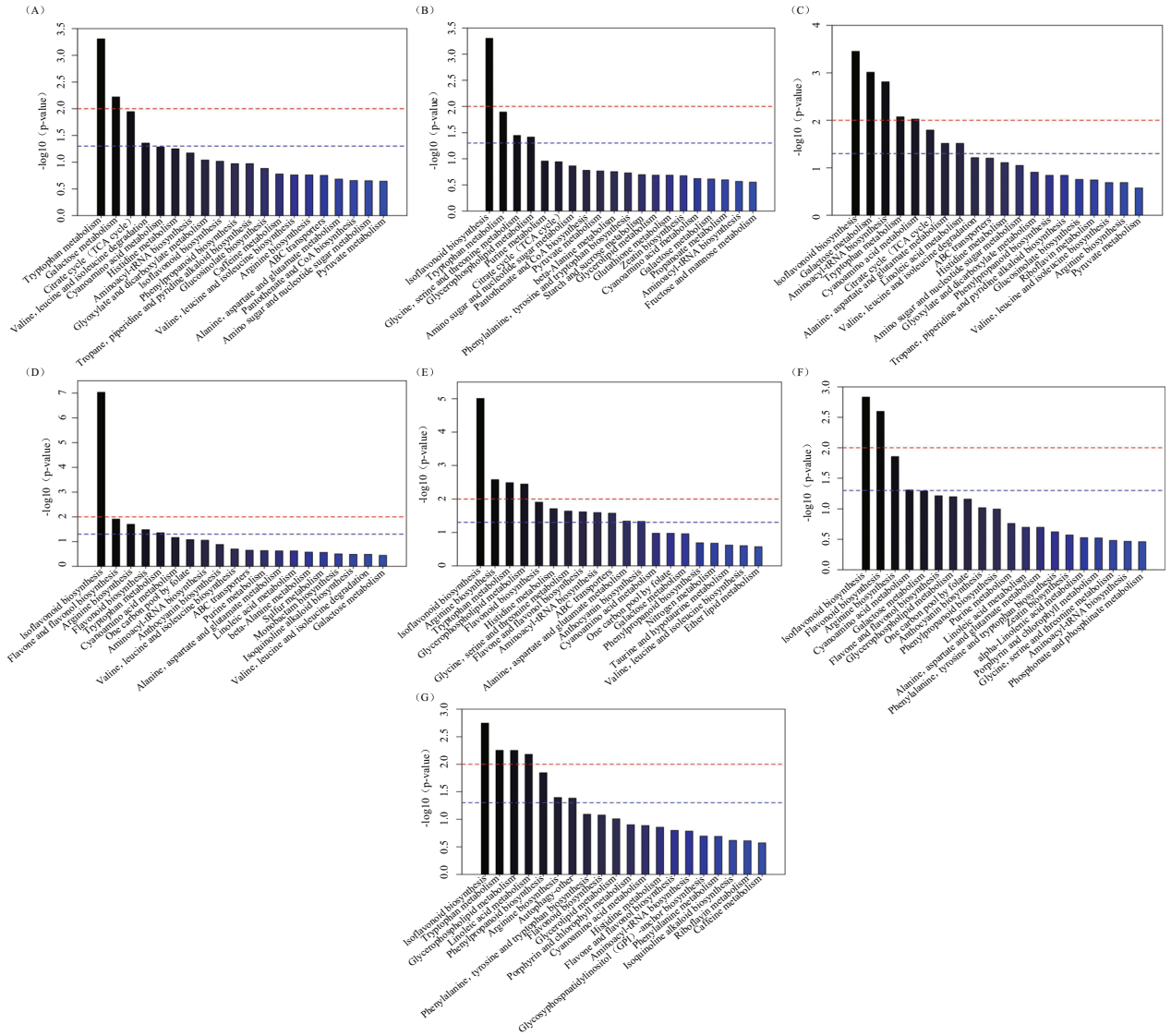

Fig. 2 Top 20 differentially enriched pathways of DEMs in each pairwise comparison. $p < 0.01$, above the red line; $p < 0.05$, between the blue line and red line. **A** DS22/SN52, **B** HH43/SN52, **C** XN20/SN52, **D** LD15/TF31, **E** LD36/TF31, **F** TD53/TF31, **G** TD67/TF31

change (FC) was calculated and the negative binomial (NB) distribution test, was used to determine the significance of differential expression. Differentially expressed genes were screened according to FC values and significance testing. The conditions for screening DEGs were FC > 1.5 or < 0.67, and $p < 0.05$.

In sample LD36/TF31 from Liaoning province, there were 4989 DEGs, of which 3032 transcripts were up-regulated and 1957 down-regulated. These were determined to be more than those of other pairwise comparisons in Liaoning. The XN20/SN52 sample of Heilongjiang province had the most DEGs among all samples within that province (2473), of which 1491 transcripts were up-regulated and 982 down-regulated (Additional file 1: Fig. S3). To better understand the association between DEMs and DEGs, KEGG enrichment analysis was carried out for each DEG pair, as shown in Additional file 1: Fig. S4.

Key metabolome and transcriptome pathway identification

Following single-dimensional and pathway analysis of the metabolome and transcriptome, we used R software to conduct Venn statistical analysis of the chosen DEMs and DEGs to exclude the possibility of gene mutation and variety specificity [56]. Because the common DEMs and DEGs obtained could be a reflection of

basic biological processes ongoing in soybeans that are driven by environmental or other factors, this analysis allowed us to further reduce the dimension and screen out shared pathways (Additional file 1: Fig. S5).

Based on the enrichment analysis, the KEGG pathways shared by DEMs and DEGs in each pairwise comparison were screened (Table 1). It was determined that not only are common biological pathways in soybeans of the two regions distinct, but pairwise comparisons revealed that soybeans from within each respective region were different. Among three pairwise comparisons of Heilongjiang soybeans, the two common pathways within the XN20/SN52 comparison, namely, galactose and alanine/aspartate/glutamate metabolism, were significantly enriched in both XN20 and SN52 groups. Among pairwise comparisons of Liaoning, flavonoid biosynthesis within LD15 and TF31, aspartate/glutamate and glycine/serine/threonine metabolism within LD36 and TF31, galactose and cyanoamino acid metabolism, and isoflavonoid biosynthesis within TD53 and TF31, and tryptophan metabolism and isoflavonoid biosynthesis within TD67 and TF31 were significantly enriched.

The bubble sorting algorithm in R was used to analyze the common enriched pathways of seven pairwise comparisons, as shown in Figs. 3 and 4. This revealed that DEGs and DEMs of related pathways were different in each pairwise comparison, even though different comparisons had the same enriched pathways. The galactose metabolism pathway involved the most pairwise comparisons. In the XN20/SN52 comparison, there were 7 DEGs within the galactose metabolism pathway; in the LD15/TF31, LD36/TF31, TD53/TF31, and TD67/TF31 comparisons, there were 15, 17, 16, 13 DEGs within the galactose metabolism pathway respectively. Correspondingly, four relevant DEMs (α-lactose, D-Gal α1->6D-Gal α1->6D-glucose, galactinol and sucrose) were identified in the XN20/SN52 comparison, along with D-Gal α1->6D-Gal α 1->6D-glucose in LD15/TF31 and TD67/TF3 comparisons, D-Gal α1->6D-Gal α1->6D-glucose and sucrose in LD36/TF31 and TD53/TF31 comparisons, respectively. Nine DEGs related to alanine/aspartic acid/glutamate metabolism were identified in the XN20/SN52 comparison, along with 13 in the LD15/TF31 comparison, 12 in the LD36/TF31 comparison, and 8 in the TD53/TF31 comparison. Correspondingly, two relevant DEMs-citric acid and L-asparagine-were identified in the XN20/SN52 comparison, along with argininosuccinate and L-glutamate in the LD36/TF31 and TD53/TF31 comparisons, and argininosuccinate in the LD15/TF31 comparison.

Hierarchical clustering of DEMs in common pathways

To further screen out potential main features and comprehensively display the relationship between soybean samples and the DEMs among each sample, we used R software to perform unsupervised hierarchical clustering on common pathway DEMs using seven pairwise comparisons [60]. Potential biomarker signatures were identified, and visual clustering based on Euclidean distance was achieved, as shown in Fig. 5.

In the DS22/SN52 comparison of soybeans from Heilongjiang, significant DEMs were citric acid and L-phenylalanine. DEM abundance of soybeans from Bayan County (SN52) was higher than that of Hailun City (DS22). The only DEM identified in the HH43/SN52 comparison was L-phenylalanine, and its abundance in Bayan County (SN52) was higher than that in Nenjiang (HH43). There were two kinds of significant DEMs identified in the XN20/SN52 comparison, one was citric and isocitric acid, the other α-lactose and sucrose. The abundance of two DEMs in Bayan County (SN52) was higher than that in Hailun City (DS22).

In the LD15/TF31 comparison of soybeans from Liaoning, significant DEMs were sulfate, secologanin, epigallocatechin, pelargonidin and D-gal alpha 1->6 D-gal alpha 1->6 D-glucose. The abundance of osteoglobulin and D-Gal alpha 1->6D-Gal alpha 1->6D-Glucose in Xinmin City (LD15) was higher than that in Jinlandian District (TF31), while sulfate, secologanin, and (-)-epigallocatechin of LD15 were less than those of TF31. In the LD36/TF31 comparison, significant DEMs were choline, L-glutamate, APC, D-gal alpha 1->6 D-gal alpha 1->6 D-glucose, sucrose, and osteoglobulin. The abundance of these metabolites in samples from Huludao City (LD36) was higher than the samples from Jinlandian District (TF31), whereas choline and L-glutamate in LD36 were less than those of TF31. In the TD67/TF31 comparison, significant DEMs were D-gal alpha 1->6 D-gal alpha 1->6 D-glucose, APC, L-phenylalanine, (-)-epigallocatechin, coenzyme A, daidzin, 6″-malonylgenistin and 4-hydroxycinnamic acid. The abundance of D-gal alpha 1->6 D-gal alpha 1->6 D-glucose, coenzyme A and APC in samples from Zhuanghe City (LD36) was higher than that the samples from Jinlandian District (TF31), while (-)-epigallocatechin, daidzin, 6″-malonylgenistin and 4-hydroxycinnamic acid of the LD36 samples were lower than those of the TF31 samples. In the TD53/TF31 comparison, significant DEMs were D-gal alpha 1->6 D-gal alpha 1->6 D-glucose, L-phenylalanine, and CMP-N-glycoloylneuraminate. D-gal alpha 1->6 D-gal alpha 1->6 D-glucose in samples from Linghai City (TD53) was more than in samples from Jinlandian District (TF31), but L-phenylalanine and CMP-N-glycolylneuraminate of TD53 were less than those of TF31. These significant DEMs establish

Table 1 Common KEGG pathways between DEMs and DEGs

Heilongjiang			**Liaoning**			
DSS22/SN52	HH43/SN52	XN20/SN52	LD15/TF31	LD36/TF31	TD53/TF31	TD67/TF31
Pyruvate metabolism	Tropane, piperidine and pyridine alkaloid biosynthesis	Galactose metabolism	Galactose metabolism	Galactose metabolism	Galactose metabolism	Galactose metabolism
Glyoxylate and dicarboxylate metabolism		Alanine, aspartate and glutamate metabolism	Alanine, aspartate and glutamate metabolism	Alanine, aspartate and glutamate metabolism	Arginine biosynthesis	Fatty acid degradation
Glucosinolate biosynthesis		Glyoxylate and dicarboxylate metabolism	Glycine, serine and threonine metabolism	Glycine, serine and threonine metabolism	Alanine, aspartate and glutamate metabolism	Glycine, serine and threonine metabolism
			Cysteine and methionine metabolism	Cysteine and methionine metabolism	Cyanoamino acid metabolism	Cysteine and methionine metabolism
			Flavonoid biosynthesis	Arginine and proline metabolism	Amino sugar and nucleotide sugar metabolism	Tyrosine metabolism
			Isoquinoline alkaloid biosynthesis	Starch and sucrose metabolism	alpha-Linolenic acid metabolism	Pantothenate and CoA biosynthesis
				Amino sugar and nucleotide sugar metabolism	Phenylpropanoid biosynthesis	Sulfur metabolism
				alpha-Linolenic acid metabolism	Isoflavonoid biosynthesis	Flavonoid biosynthesis
				Glyoxylate and dicarboxylate metabolism	Isoquinoline alkaloid biosynthesis	Isoflavonoid biosynthesis
				Phenylpropanoid biosynthesis		Isoquinoline alkaloid biosynthesis
				Flavonoid biosynthesis		Glucosinolate biosynthesis
						Biosynthesis of unsaturated fatty acids

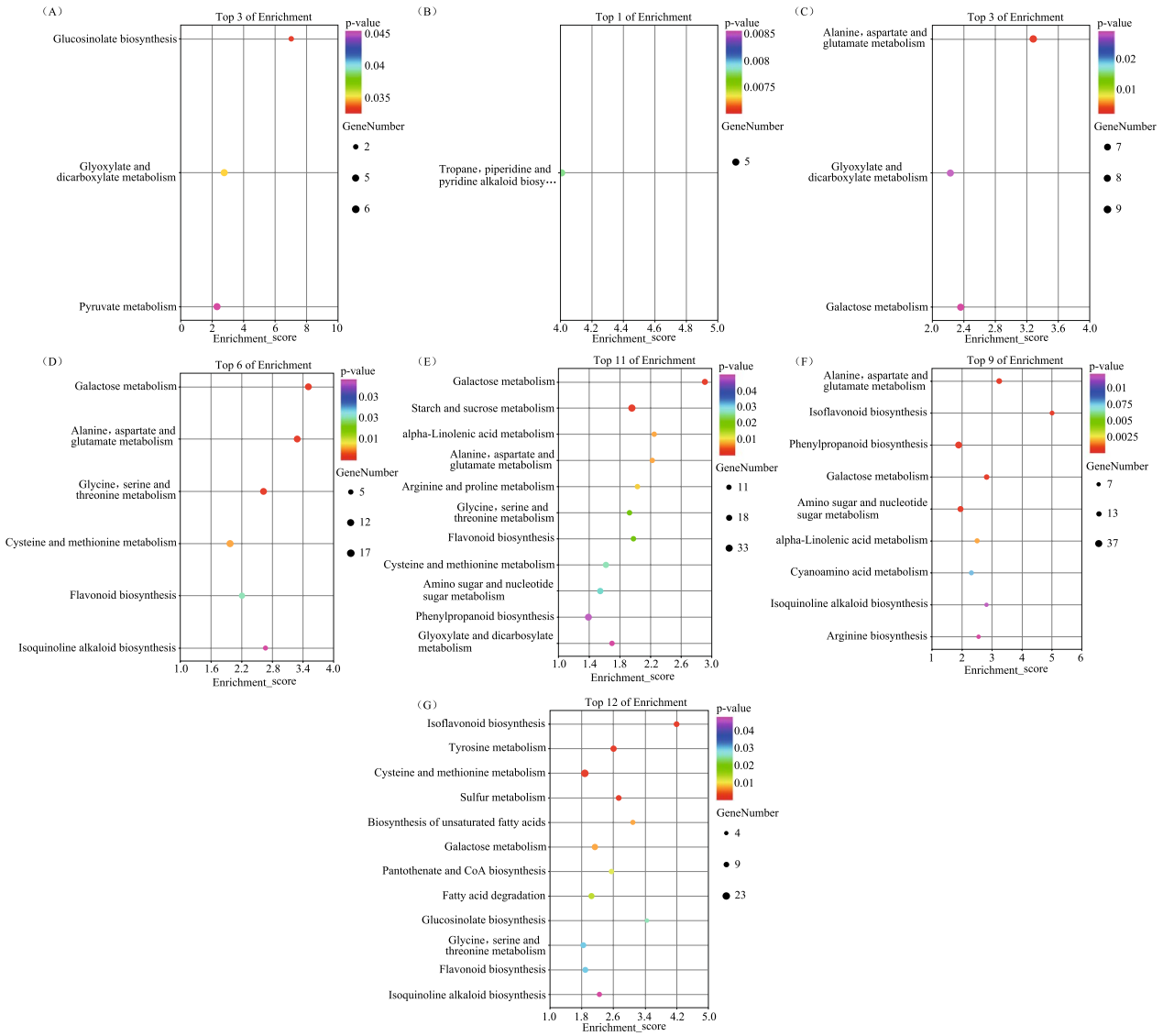

Fig. 3 Enriched KEGG pathways of DEGs. **A** DS22/SN52, **B** HH43/SN52, **C** XN20/SN52, **D** LD15/TF31, **E** LD36/TF31, **F** TD53/TF31, **G** TD67/TF31

a basis for further exploring the use of these metabolites as potential soybean biomarkers in northern China.

Integration of metabolomic and transcriptomic analysis

Based on the transcriptomic and metabonomic analysis, corresponding mRNAs and metabolites relevant to each other were comprehensively screened in the different environments. The pathway-based method was used to integrate and analyze the relationship between gene transcription and metabolism so as to comprehensively understand the biological pathways which respond to environmental changes during soybean growth in each region, and to identify possible molecular signatures (i.e., biomarkers) which are unique to each region. Based on Spearman correlation coefficients of DEGs and DEMs in the shared pathways, the relationship between variables was revealed using correlation network analysis, as shown in Additional file 1: Fig. S6.

In order to further characterize differences between significant DEGs and DEMs among soybeans from various regions, Spearman correlation coefficient analysis was applied [61]. Hierarchical clustering (Fig. 6) was conducted through R, which showed that there were 11 very strong and 54 strong correlations between DEMs and DEGs in the TD53/TF31 comparison. There were 17 very strong and 114 strong correlations in the LD36/TF31 comparison, 8 very strong and 39 strong

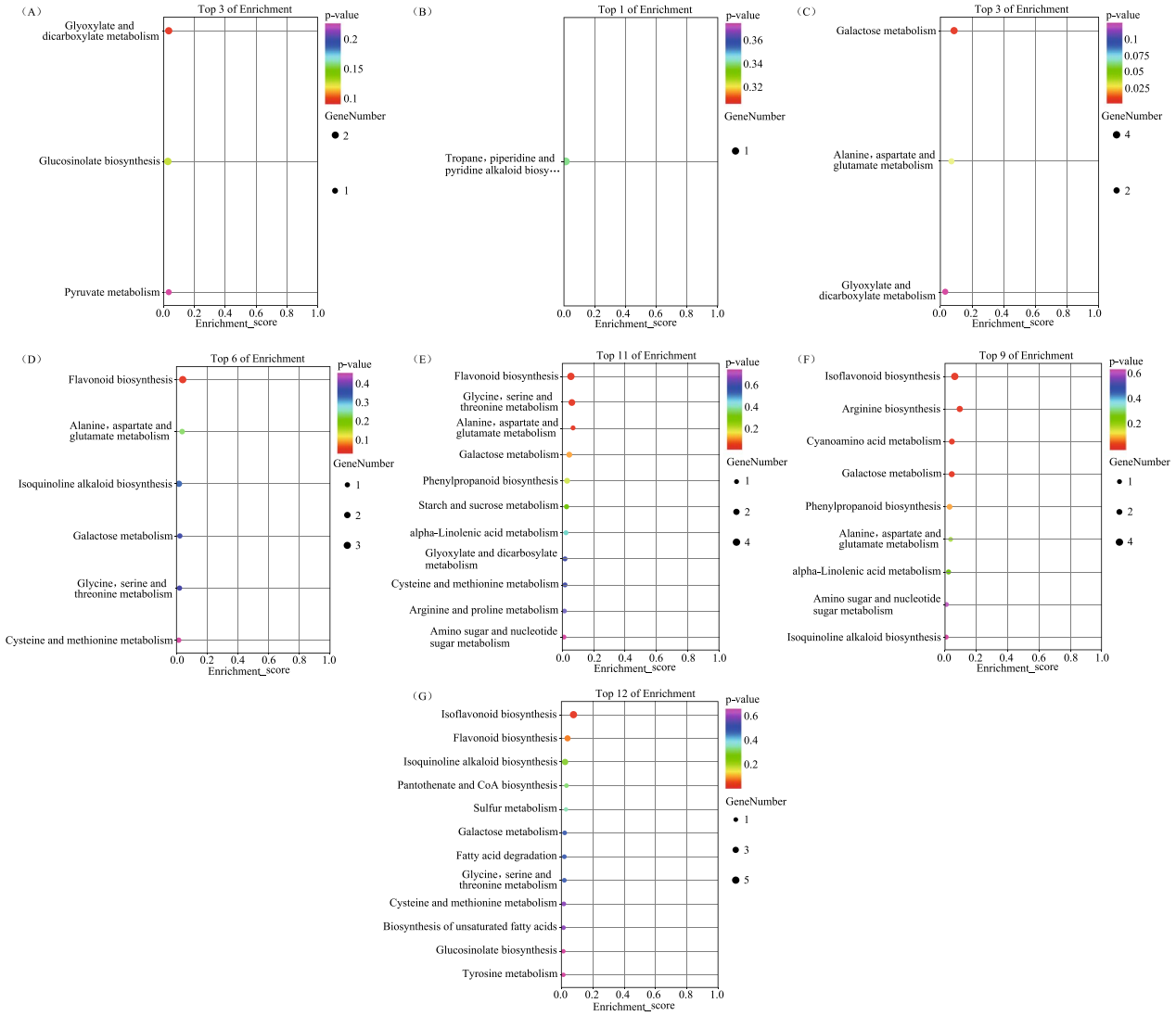

Fig. 4 Enriched KEGG pathways of DEMs. The ordinate is the name of a metabolic pathway, the abscissa is the Rich factor (number of statistically significant DEMs/total number of metabolites in this pathway). The larger the Rich factor, the greater the enrichment degree. The color from green to red indicates that the *p*-value decreases in turn. The larger the dot, the more metabolites enriched on this pathway. **A** DS22/SN52, **B** HH43/SN52, **C** XN20/SN52, **D** LD15/TF31, **E** LD36/TF31, **F** TD53/TF31, **G** TD67/TF31

correlations in the LD15/TF31 comparison, 11 very strong and 94 strong correlations in the TD67/TF31 comparison. There were two very strong and 44 strong correlations in the XN20/SN52 comparison, one very strong and 16 strong correlations in the DS22/SN52 comparison. The exact biological interpretation of these correlations is still unclear, however, and the DEMs and DEGs screened need to be further verified under various environmental conditions before any of them can be assigned as bona fide biomarkers. However, it is clear from the correlation between DEMs and DEGs that differing environments in different geographical regions impacted DEM abundance [62].

Core DEMs

We identified 42 significant DEMs that have the potential to become novel biomarker signatures capable of distinguishing soybeans from different production areas (Table 2). Venn statistical analysis was conducted on these 42 DEMs to find representative core DEMs as shown in Additional file 1: Fig. S7. In Heilongjiang province, core DEMs were citric acid, isocitric acid and L-phenylalanine (Additional file 1: Table S3). In Liaoning

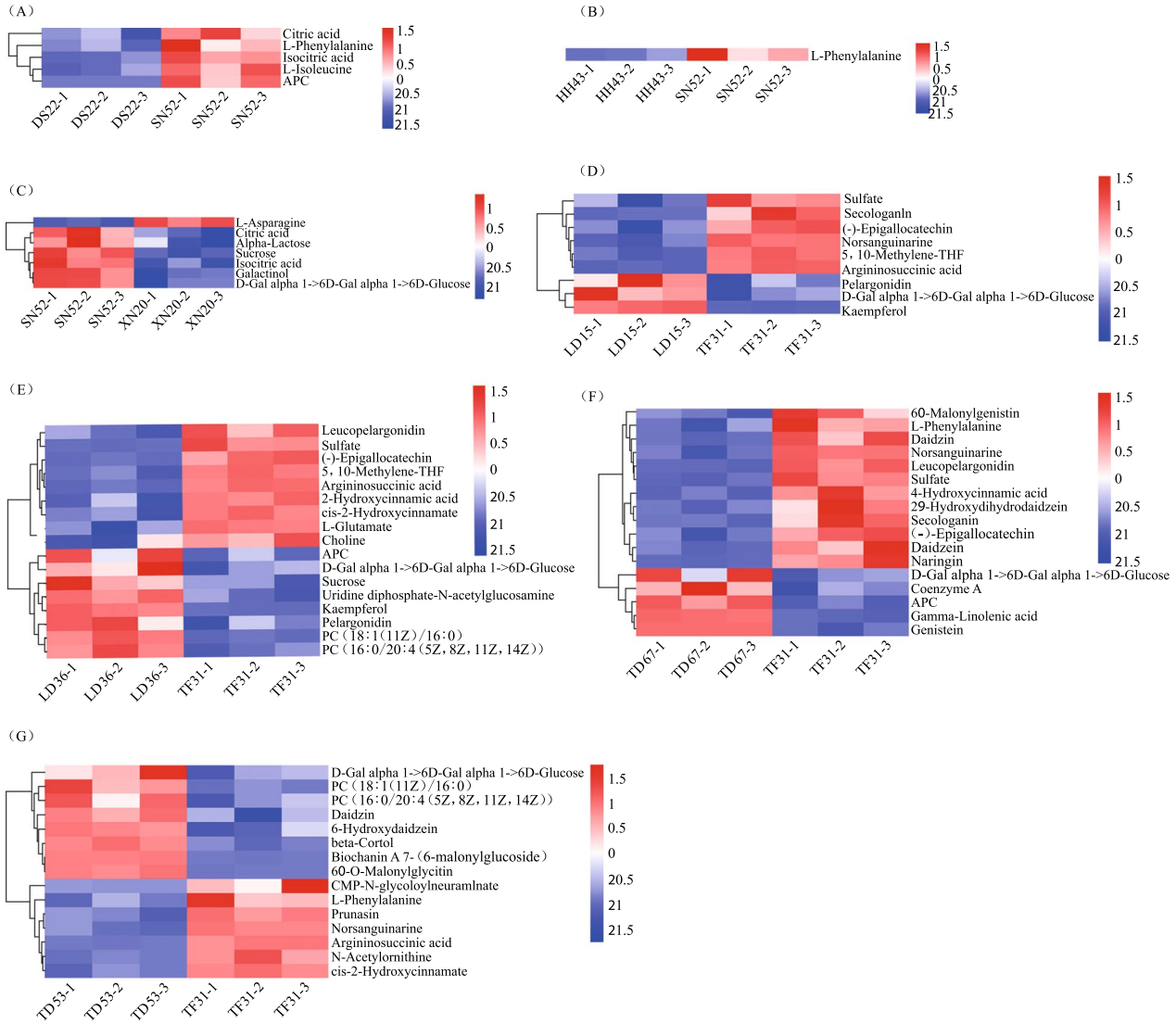

Fig. 5 Heatmap of DEMs in pairwise comparisons. The abscissa represents the sample name, and the ordinate represents DEM. The left branch shows the clustering of DEMs, and the color from blue to red represents the expression abundance of metabolites from low to high; the more intense red and blue colors represent the magnitude of DEM expression. **A** DS22/SN52, **B** HH43/SN52, **C** XN20/SN52, **D** LD15/TF31, **E** LD36/TF31, **F** TD67/TF31, **G** TD53/TF31

province, the most core DEM was D-gal alpha1->6 D-gal alpha 1->6 D-glucose, followed by sulfate, (-)-epigallocatechin, argininosuccinic acid and norsanguinarine (Additional file 1: Table S4).

Partial least squares discriminant analysis of core DEMs

To determine whether D-Gal alpha 1->6D-Gal alpha 1->6D-Glucose, citric acid, isocitric acid, L-phenylalanine, sulfate, (-)-epigallocathin, argininosuccinic acid and norsanguinarine can be used as classification indicators to distinguish soybeans from different production areas, PLS-DA analysis was used on these DEMs as shown in Additional file 1: Fig. S8. In the PLS-DA model, R2X and R2Y represent the percentage of X and Y matrix information respectively. R2X (cum) = 90.9%, indicating that the three principal prediction components of the model can explain 90.9% of the X variable; R2Y (cum) = 100%, indicating that the three principal prediction components of the model can explain 100% of the Y variable. Q2 indicates that the prediction ability of the evaluation model was obtained through cross validation. Here, Q2 = 0.944, indicating that the prediction ability of PLS-DA for soybean samples in the nine production areas of the two provinces (Bei'an, Nenjiang, Hailun, Bayan, Zhuanghe,

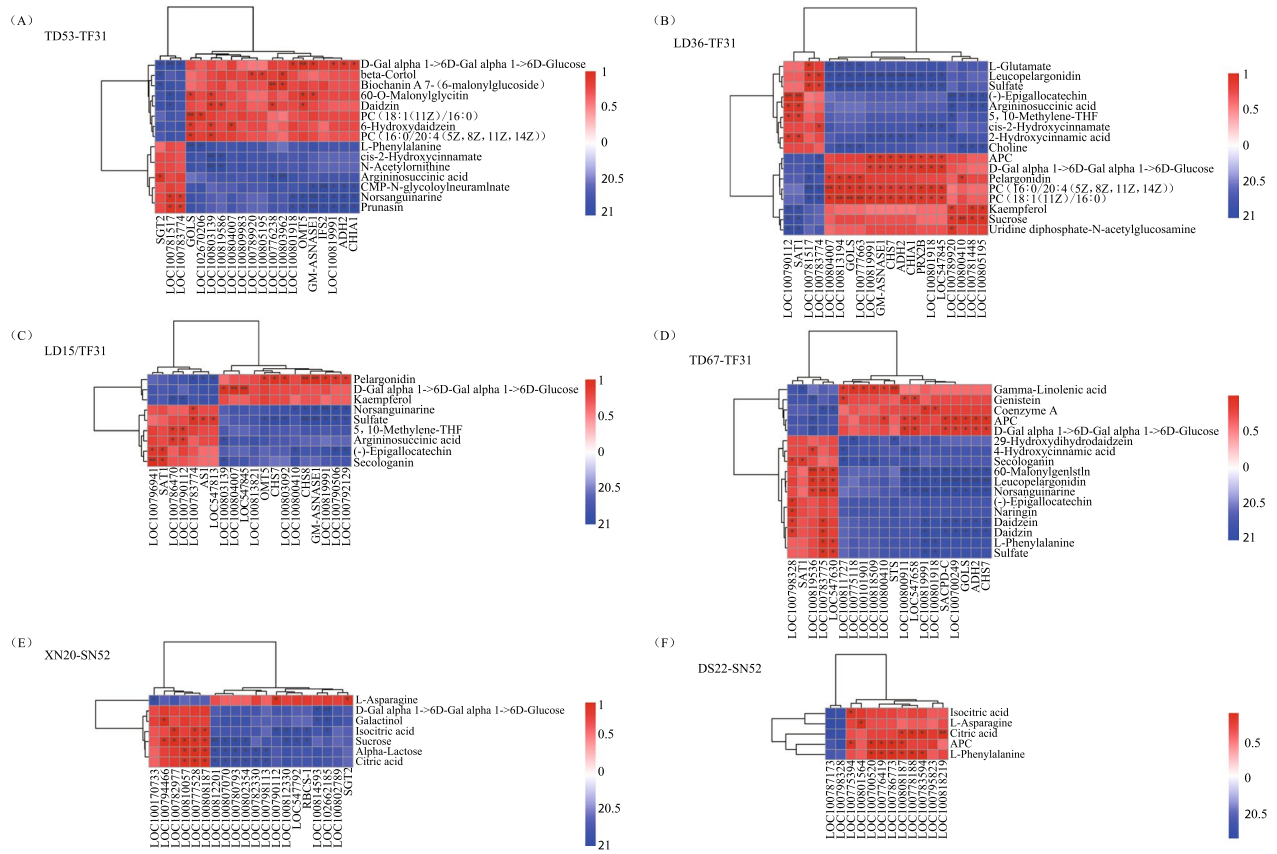

Fig. 6 Correlation heatmap of DEMs and DEGs. The left branch shows the clustering of DEMs, and the upper branch shows the clustering of DEGs. The correlation coefficient r is expressed in color; $r > 0$ represents positive correlation, expressed in red, while $r < 0$ represents negative correlation, expressed in blue. The darker the color is, the stronger the correlation is. The *p*-value reflects the significant level of correlation. $^{*}p < 0.1$, $^{**}p < 0.05$. **A** TD53/TF31, **B** LD36/TF31, **C** LD15/TF31, **D** TD67/TF31, **E** XN20/SN52, **F** DS22/SN52

Huludao, Jinlandian, Xinmin and Linghai) was 94.4%. Additional file 1: Fig. S8 shows that all soybean samples were divided into different regions, and soybeans from the same production region had aggregated R2Z/Y values. These findings suggest that the seven classification indexes contained enough information to accurately identify and distinguish soybean samples from nine production areas.

Discussion

In this study, we found that an unbiased metabolomics approach using LC–MS provides a comprehensive comparison of metabolic characteristics among soybean samples obtained from different geographical regions in China. These data yield key characteristics about the soybeans in high dimension. A total of 8283 metabolites were detected by LC–MS. Of these, 42 significant DEMs were identified by the combination of single-dimensional and multi-dimensional statistical methods with pathway-based integrated analysis. These DEMs were determined to mainly be comprised of flavonoids, isoflavonids, organooxygen compounds, carboxylic acids and derivatives. Among them, eight DEMs belong to the family of carboxylic acid derivatives, followed by seven isoflavonoids and seven organooxygen compounds.

The composition of metabolites in agricultural products not only depends on the underlying genetics of the various species tested but also is heavily influenced by the natural environment of their cultivation [62]. This is reflected in the significant differences in the metabolite composition of soybeans from different production areas in the present study. Plant metabolism is also highly responsive to climate (temperature, precipitation) and geographical location (altitude, longitude and latitude) [63]. We analyzed the climate and geographical characteristics of nine regions, including average temperature, rainfall, sunshine time, etc. Daily sunshine duration can affect photosynthesis rates in soybean plants, thus affecting carbohydrate metabolism [64], formation of flavonoids [65] and other metabolites. Annual average temperature has a substantial impact on lipid metabolism. In lower temperatures, changes in

Table 2 Significant DEMs

DEM	Type of compounds
(-)-Epigallocatechin	Flavonoids
2-Hydroxycinnamic acid	Cinnamic acids and derivatives
2 -Hydroxydihydrodaidzein	Isoflavonoids
4-Hydroxycinnamic acid	Cinnamic acids and derivatives
5,10-Methylene-THF	Pteridines and derivatives
6-Hydroxydaidzein	Isoflavonoids
6 -Malonylgenistin	Isoflavonoids
6 -*O*-Malonylglycitin	Isoflavonoids
Alpha-Lactose	Organooxygen compounds
APC	Organooxygen compounds
Argininosuccinic acid	Carboxylic acids and derivatives
beta-Cortol	Organooxygen compounds
Biochanin A 7-(6-malonylglucoside)	Benzofurans
Choline	Organonitrogen compounds
cis-2-Hydroxycinnamate	Cinnamic acids and derivatives
Citric acid	Carboxylic acids and derivatives
CMP-*N*-glycoloylneuraminate	Pyrimidine nucleotides
Coenzyme A	Purine nucleotides
Daidzein	Isoflavonoids
Daidzin	Isoflavonoids
D-Galalpha 1->6D-Gal alpha 1->6D-Glucose	Organooxygen compounds
Galactinol	Organooxygen compounds
Gamma-Linolenic acid	Fatty Acyls
Genistein	Isoflavonoids
Isocitric acid	Carboxylic acids and derivatives
Kaempferol	Flavonoids
L-Asparagine	Carboxylic acids and derivatives
Leucopelargonidin	Flavonoids
L-Glutamate	Carboxylic acids and derivatives
L-Isoleucine	Carboxylic acids and derivatives
L-Phenylalanine	Carboxylic acids and derivatives
N-Acetylornithine	Carboxylic acids and derivatives
Naringin	Flavonoids
Norsanguinarine	Quinolines and derivatives
PC (16:0/20:4(5Z,8Z,11Z,14Z))	Glycerophospholipids
PC (18:1(11Z)/16:0)	Glycerophospholipids
Pelargonidin	Flavonoids
Prunasin	Organooxygen compounds
Secologanin	Prenol lipids
Sucrose	Organooxygen compounds
Sulfate	Non-metal oxoanionic compounds
Uridine diphosphate-*N*-acetylglucosamine	Pyrimidine nucleotides

enzyme activity lead to lipid accumulation [65]. Average annual precipitation and soil composition are factors which can impact the accumulation of carboxylic acids, derivatives and non-metallic oxygen compounds.

The annual average temperature and length of summer in Heilongjiang Province is lower than in Liaoning Province. These differences likely explain our finding that L-phynylalanine, citric acid and isocitric acid metabolites

were higher in soybeans from Heilongjiang as compared with those from Liaoning. The annual average precipitation and sunshine days are higher in Heilongjiang as compared with those from Liaoning. Abundant precipitation greatly increases soil organic matter accumulation and transformation, and abundant sunlight during soybean growth period likely promoted the increase of (-)-epigallocatechin, D-Gal α1->6 D-Gal α1->6 D-glucose and other metabolites in Heilongjiang soybeans. Organic matter accumulation and humus content are impacted by differences in terroir of the various soybean production areas examined in this study. Through the comparison of DEM heat maps, we found that carboxylic acids and their derivatives in Heilongjiang samples were higher than in Liaoning samples. Heilongjiang has low temperature, abundant precipitation and robust black soil aggregation. This favorable environment has greatly facilitated the material exchange of various organic matters in the "soil-soybean plant" interaction, and promoted high quality soybean production.

The overall cold temperate climate of Heilongjiang province is in sharp contrast to the temperate monsoon climate of Liaoning. Due to differences in longitude and latitude of different regions within each province, the average sunshine, precipitation and climate in these regions are also different, in turn promoting the increase or decrease of various metabolites. Among all environmental factors, the most important is geographical location (i.e. latitude, longitude), followed by annual average temperature, average precipitation, and soil organic matter composition [66]. Soybeans must be planted carefully and with appropriate timing so as to maximize overall ventilation and light conditions. In general, large seeds need more water and are suitable for planting in areas with sufficient rainfall, while small seeds need less water and are mostly planted in arid areas. Soybean metabolite differences are not only related to soil and climatic conditions, but also to soybean varieties and agricultural practices. We screened metabolites directly related to geographical origin in this study, which was combined with transcriptome sequencing and characterization to reduce genotype interference. However, differences in some metabolites among the various soybean samples may be due to the different varieties grown in the different production areas. In future studies, both soybean varieties and agricultural practices should be considered when determining the geographical origin of the samples that will be studied.

In the comparison study of samples DS22, SN52, LD15, TF31, LD36, TF31, LD53, TF31, LD67, TF31, XN20, and SN52, it was found that there are core differential substances present in DS22 and SN52 samples, including key metabolites in amino acid and the citric acid cycle. The biological pathways and regulatory genes involved in these differences are as follows: Differential metabolites related to the citric acid cycle include Citric acid and Isocitric acid, which are essential components in the respiratory process, providing precursors for energy production and various biosynthetic pathways. Differential genes may be involved in regulating the enzymes of the TCA cycle, impacting the production and conversion of citric acid and isocitric acid. In terms of amino acid biosynthesis, L-Isoleucine and L-Phenylalanine are essential amino acids for protein biosynthesis and serve as precursors for certain secondary metabolites, such as alkaloids and flavonoids. The biosynthesis of these amino acids may be regulated by core differential genes.

From a genetic perspective, core differential genes such as LOC100775394, LOC100776419, and LOC100778188 may be involved in regulating relevant metabolic pathways. L-Phenylalanine, as a core differential substance, appears in the HH43 and SN52 samples, indicating that genes related to phenylalanine biosynthesis pathways may have different expression patterns between these two samples. These genes (such as LOC100306108, LOC100780806, LOC100785449, LOC100814593, LOC547792) may have different allelic versions, or their expression levels may be regulated, influencing the synthesis and accumulation of phenylalanine.

The core differential substances between LD15 and TF31 samples may be related to the growth environment of the samples, such as soil composition and climatic conditions. These environmental factors can influence the type and quantity of final metabolites by affecting the expression or activity of plant endogenous metabolic pathways. For example, (-)-Epigallocatechin is a flavonoid compound found in high concentrations in green tea, and Kaempferol is a widely distributed flavonoid found in plants. Norsanguinarine and Pelargonidin are respectively alkaloids and anthocyanins, and their biosynthesis may be induced or inhibited by environmental factors. Core differential genes such as AS1, CHS7, CHS8, and CHS9 may participate in the biosynthesis, regulation, and response to environmental changes of the above-mentioned metabolites.

The core differential substances between LD36 and TF31 samples indicate significant differences in lipid metabolism and polyphenol synthesis pathways. PC (16:0/20:4(5Z,8Z,11Z,14Z)) and PC (18:1(11Z)/16:0) are specific molecular species of phospholipids, playing important roles in cellular membrane structure and function, signal transduction, and lipid storage. The synthesis and metabolism of PC involve various enzymes, such as PCYT and PLA2, and the activity and regulation of these enzymes may be related to differential gene expression. Pelargonidin is an anthocyanin, a class of plant secondary

metabolites with antioxidant properties, directly related to flower coloration. Its biosynthesis pathway involves the participation of multiple enzymes, such as CHS and CHI. Sucrose, Sulfate, and Uridine diphosphate-N-acetylglucosamine are key molecules in the metabolic process, participating in carbohydrate metabolism, sulfur assimilation and transport, and glycosyl transfer reactions, respectively. Core differential genes such as AS1, COI1 may be involved in the above metabolic pathways and biological processes.

The core differential substances and differential genes between LD53 and TF31 samples indicate significant biochemical differences between these two samples, which may result from the interaction of environmental influences and genetic factors. Core differential substances such as Argininosuccinic acid are important intermediates in the urea cycle, related to the detoxification and excretion of ammonia; beta-Cortol, Biochanin A 7-(6-malonylglucoside), cis-2-Hydroxycinnamate, Daidzin, and Norsanguinarine are plant secondary metabolites related to plant defense mechanisms, pigment formation, and interaction with the environment. *N*-Acetylornithine is involved in amino acid conversion and metabolism intermediates; PC (16:0/20:4(5Z,8Z,11Z,14Z)) and P C(18:1(11Z)/16:0) are constituents of membrane lipids, affecting cellular membrane structure and function; Prunasin is a cyanogenic compound, possibly involved in plant natural defense mechanisms. Core differential genes such as 4CL4, ADH2, AOC1, CHIA1, CYP93A1, GDH2, GS, GS1GAMMA2, GM-ASNASE1, GMPAL2.1, GMPAL2.3, and GOLS may be involved in the synthesis, regulation, and response to environmental changes of the above-mentioned metabolites.

The core differential substances between LD67 and TF31 samples indicate significant differences in biochemical composition. These differences may reflect the physiological state, growth conditions, environmental stress adaptation, and metabolic pathway characteristics of the two samples. Core differential substances, such as (-)-Epigallocatechin, Daidzein, Genistein, and other flavonoid and isoflavone compounds, are commonly associated with antioxidant, defense responses, and signal transduction. 4-Hydroxycinnamic acid participates in the biosynthesis of phenolic compounds and has antioxidant properties. Gamma-Linolenic acid is a polyunsaturated fatty acid that is crucial for regulating inflammatory responses and maintaining cellular membrane function. Core differential genes such as ADH2 play a role in the alcohol dehydrogenase-derived products process in estimating yeast. CHS7, CHS8, and CHS9 are related to the biosynthesis of flavonoid compounds. SACPD-C participates in the unsaturation of fatty acids. Coenzyme A is an important cofactor that functions in various biological reactions, such as fatty acid and amino acid metabolism. OAS-TL2 and OAS-TL3 participate in sulfur metabolism and protein synthesis.

The core differential substances between XN20 and SN52 samples indicate significant metabolic differences, which may be the result of multiple factors interacting. Core differential substances such as Alpha-Lactose are a form of lactose found in dairy products, associated with energy supply and carbohydrate metabolism; Citric acid and Isocitric acid are key metabolites in the tricarboxylic acid cycle, crucial for energy production and metabolic regulation; D-Gal α-1->6D-Gal α-1->6D-Glucose is a compound containing galactose and glucose residues, possibly related to sugar transport and signaling; Galactinol is a plant sugar alcohol, serving as an osmoprotectant in response to stress conditions such as drought. L-Asparagine is a nonessential amino acid playing a role in nitrogen transport and storage. Sucrose is the most common non-structural carbohydrate in plants, used for energy storage and transport. Core differential gene RBCS-1 is typically associated with photosynthesis in plants, participating in the carbon fixation process.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13765-024-00882-x.

Additional file 1: Fig. S1: **a** PCA of metabolites in soybeans from different regions; **b** PCA of gene transcripts in soybeans from different regions. **Fig. S2**: The number of DEMs in each group pair. **Fig. S3**: Number of DEGs of each pairwise comparison. **Fig. S4**: KEGG pathway annotation of DEGs in each pairwise comparison. **A** DS22/SN52, **B** HH43/SN52, **C** XN20/SN52, **D** LD15/TF31, **E** LD36/TF31, **F** TD53/TF31, **G** TD67/TF31. **Fig. S5**: Venn diagram of the common DEMs and DEGs in **a** Heilongjiang provinces and **b** Liaoning provinces. **Fig. S6**: Connection network between metabolites and genes inferred from metabolomic and transcriptomic analysis. Correlation coefficient $|r| \geq 0.5$, $p < 0.05$. **A** DS22/SN52, **B** XN20/SN52, **C** LD15/TF31, **D** LD36/TF31, **E** TD53/TF31, **F** TD67/TF31. **Fig. S7**: **a** Upset diagram of core DEMs in Heilongjiang; **b** Upset diagram of core DEMs in Liaoning. **Fig. S8**: Scatterplot of PLS-DA model for soybean samples from different sources. **Table S1**: The types and quantities of metabolites contained in various soybean samples. **Table S2**: The total number of detectable expressed genes in soybeans in each region; **Table S3**: Core DEMs among pairwise comparisons in Heilongjiang; **Table S4**: Core DEMs among pairwise comparisons in Liaoning.

Acknowledgements
We sincerely thank the local Academy of Agricultural Sciences and related research institutions where the samples are located for providing soybeans for this study.

Author contributions
The authors confirm contribution to the paper as follows: study conception and design: WJ. data collection: WJ, ZQ. analysis and interpretation of results: ZQ, WC. draft manuscript preparation: WJ, ZA. All authors reviewed the results and approved the final version of the manuscript.

Funding
This research received no external funding.

Availability of data and materials
All data generated during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 10 November 2023 Accepted: 3 March 2024
Published online: 16 March 2024

References

1. Zappi A, Melucci D, Scaramagli S et al (2018) Botanical traceability of unifloral honeys by chemometrics based on head-space gas chromatography. Eur Food Res Technol 244(12):2149–2157
2. Jamet JP, Chaumet JM (2016) Soybean in China: adaptating to the liberalization. Ocl 23(6):D604
3. Aung MM, Chang YS (2014) Traceability in a food supply chain: Safety and quality perspectives. Food Control 39:172–184
4. Jiang ZQ (2018) Research progress on traceability of grain origin produced by mineral element fingerprint analysis technology. Farm Products Process 5:70–71
5. Zhao S, Zhao Y (2021) Application and preparation progress of stable isotope reference materials in traceability of agricultural products. Crit Rev Anal Chem 51(8):742–753
6. Zhang Y, Wang D, Li X (2018) Research progress on origin tracing of agricultural products based on near infrared spectroscopy. J Food Saf Qual 9:6161–6166
7. Sheng CD, Yu JH, Qing LH et al (2020) Geographical specificity of fatty acid and multi-element fingerprints of soybean in northern China. Qual Assurance Saf Crops Foods 12(3):126–139
8. Wang ZC, Yan Y, Nisar T et al (2019) Multivariate statistical analysis combined with e-nose and e-tongue assays simplifies the tracing of geographical origins of Lycium ruthenicum Murray grown in China. Food Control 98:457–464
9. Jewett MC, Hofmann G, Nielsen J (2006) Fungal metabolite analysis in genomics and phenomics. Curr Opin Biotechnol 17(2):191–197
10. Khalid N, Aqeel M, Noman A (2019) System biology of metal tolerance in plants: An integrated view of genomics, transcriptomics, metabolomics, and phenomics. Plant Metall Funct Omics 2019:107–144
11. Singh S et al (2016) Heavy metal tolerance in plants: role of transcriptomics, proteomics, metabolomics, and ionomics. Front Plant Sci 6:1143
12. Tiedge K et al (2022) Comparative transcriptomics and metabolomics reveal specialized metabolite drought stress responses in switchgrass (*Panicum virgatum*). New Phytol 236(4):1393–1408
13. Fiehn O (2002) Metabolomics the link between genotypes and phenotypes. Funct Genomics 2002:155–171
14. Severin AJ, Woody JL, Bolon YT et al (2010) RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. BMC Plant Biol 10(1):1–16
15. Xiao R, Ma Y, Zhang D et al (2018) Discrimination of conventional and organic rice using untargeted LC–MS-based metabolomics. J Cereal Sci 82:73–81
16. Gonzalez-Covarrubias V, Eduardo M-M, del Bosque-Plata L (2022) The potential of metabolomics in biomedical applications. Metabolites 12(2):194
17. Mehari B, Redi-Abshiro M, Chandravanshi BS et al (2019) GC?MS profiling of fatty acids in green coffee (*Coffea arabica* L.) beans and chemometric modeling for tracing geographical origins from Ethiopia. J Sci Food Agric 99(8):3811–3823
18. Zhang X, Liu Y, Li Y et al (2017) Identification of the geographical origins of sea cucumber (*Apostichopus japonicus*) in northern China by using stable isotope ratios and fatty acid profiles. Food Chem 218:269–276
19. Rubab S, Rizwani GH, Bahadur S et al (2020) Determination of the GC?MS analysis of seed oil and assessment of pharmacokinetics of leaf extract of *Camellia sinensis* L. J King Saud Univ Sci 32(7):3138–3144
20. Chen C-J et al (2022) Recent advances in LC–MS based metabolomics for clinical biomarker discovery. Mass Spectromet Rev 2022:21785
21. Shen S et al (2023) Metabolomics-centered mining of plant metabolic diversity and function: past decade and future perspectives. Mol Plant 16(1):43–63
22. Jing J, Shi Y, Zhang Q et al (2017) Prediction of Chinese green tea ranking by metabolite profiling using ultra-performance liquid chromatography-quadrupole time-of-flight mass spectrometry (UPLC-Q-TOF/MS). Food Chem 221:311–316
23. Lee JE, Lee BJ, Chung JO et al (2015) Metabolomic unveiling of a diverse range of green tea (*Camellia sinensis*) metabolites dependent on geography. Food Chem 174:452–459
24. Yun DY, Kang YG, Kim EH et al (2018) Metabolomics approach for understanding geographical dependence of soybean leaf metabolome. Food Res Int 106:842–852
25. Wang L, Liu L, Ma Y et al (2018) Transcriptome profilling analysis characterized the gene expression patterns responded to combined drought and heat stresses in soybean. Comput Biol Chem 77:413–429
26. Zhang Z et al (2022) Integrated metabolomics and transcriptomics analyses reveal the metabolic differences and molecular basis of nutritional quality in landraces and cultivated rice. Metabolites 12(5):384
27. Huang W et al (2022) Metabolomics and transcriptomics analysis of vitro growth in pitaya plantlets with different LED Light spectra treatment. Ind Crops Prod 186:115237
28. Zhu Z et al (2023) Transcription and metabolic profiling analysis of three discolorations in a day of hibiscus mutabilis. Biology 12(8):1115
29. Nguyen HD, Kim M-S (2022) The protective effects of curcumin on metabolic syndrome and its components: in-silico analysis for genes, transcription factors, and microRNAs involved. Arch Biochem Biophys 727:109326
30. Gong L et al (2022) Prediction of potential distribution of soybean in the frigid region in China with MaxEnt modeling. Ecol Inform 72:101834
31. Sheng CD et al (2020) Geographical specificity of fatty acid and multi-element fingerprints of soybean in northern China. Qual Assurance Saf Crops Foods 12(3):126–139
32. Nawaz MA et al (2020) Korean wild soybeans (*Glycine soja* Sieb & Zucc.): geographic distribution and germplasm conservation. Agronomy 10(2):214
33. Sachar S, Kumar A (2021) Survey of feature extraction and classification techniques to identify plant through leaves. Expert Syst Appl 167:114181
34. Zhang J et al (2021) Taxonomic compositions and co-occurrence relationships of protists in bulk soil and rhizosphere of soybean fields in different regions of China. Front Microbiol 12:738129
35. Yin L et al (2020) Optimizing feature selection of individual crop types for improved crop mapping. Remote Sens 12(1):162
36. Xiong F et al (2021) Non-target metabolomics revealed the differences between *Rh. tanguticum* plants growing under canopy and open habitats. BMC Plant Biol 21(1):1–13
37. Xian Y, Liu G, Yao H (2022) Predicting the current and future distributions of major food crop designated geographical indications (GIs) in China under climate change. Geocarto Int 37(25):8148–8171
38. Lucas KRG (2021) Using the available indicators of potential biodiversity damage for Life Cycle Assessment on soybean crop according to Brazilian ecoregions. Ecol Indic 127:107809
39. Chotekajorn A et al (2021) Evaluation of seed amino acid content and its correlation network analysis in wild soybean (*Glycine soja*) germplasm in Japan. Plant Genet Resour 19(1):35–43
40. Hu Y et al (2022) Sexual compatibility of transgenic soybean and different wild soybean populations. J Integr Agric 21(1):36–48
41. Saleem A et al (2021) A genome-wide genetic diversity scan reveals multiple signatures of selection in a European soybean collection compared to Chinese collections of wild and cultivated soybean accessions. Front Plant Sci 12:631767
42. Azizah FN et al (2023) Detection of metabolites in rhizosphere of soybean under different status of soil potassium. Soil Sci Plant Nutr 69(2):69–77

43. Liu Y et al (2022) The interrelationship between latitudinal differences and metabolic differences in the natural distribution area of Tilia amurensis Rupr. Forests 13(9):1507
44. Kim M et al (2022) RNA-seq gene profiling reveals transcriptional changes in the late phase during compatible interaction between a Korean soybean cultivar (*Glycine max* cv. Kwangan) and pseudomonas syringae pv. syringae B728a. Plant Pathol J 38(6):603
45. Durmanov A et al (2023) Sustainable growth of greenhouses: investigating key enablers and impacts. Emerg Sci J 7(5):1674–1690
46. Suseno BD (2023) Role of the magnitude of digital adaptability in sustainability of food and beverage small enterprises competitiveness. HighTech Innov J 4(2):270–282
47. Kassymbek R et al (2023) Optimization of the extrusion process in the production of compound feeds for dairy cows. Emerg Sci J 7:1574–1587
48. Yang Y et al (2022) Drought risk assessment of millet and its dynamic evolution characteristics: a case study of Liaoning Province, China. Ecol Indic 143:109407
49. Li D et al (2022) Spatial evolution of cultivated land in the Heilongjiang Province in China from 1980 to 2015. Environ Monit Assess 194(6):444
50. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120
51. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357–360
52. Roberts A, Trapnell C, Donaghey J et al (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 12(3):1–14
53. Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515
54. Putri GH et al (2022) Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics 38(10):2943–2945
55. Anders S, Huber W (2012) Differential expression of RNA-Seq data at the gene level-the DESeq package. European Molecular Biology Laboratory (EMBL), 10: f1000research, Heidelberg, Germany
56. Fouts DE, Szpakowski S, Purushe J et al (2012) Next generation sequencing to define prokaryotic and fungal diversity in the bovine rumen. PLoS ONE 7(11):e48289
57. Kanehisa M, Araki M, Goto S et al (2007) KEGG for linking genomes to life and the environment. Nucl Acids Res 36(suppl):D480–D484
58. Eriksson L, Byrne T, Johansson E et al (2013) Multi-and megavariate data analysis basic principles and applications. Umetrics Academy
59. Kanehisa M et al (2023) KEGG for taxonomy-based analysis of pathways and genomes. Nucl Acids Res 51(D1):D587–D592
60. Yang H et al (2021) Integrative analyses of metabolome and transcriptome reveals metabolomic variations and candidate genes involved in sweet cherry (*Prunus avium* L.) fruit quality during development and ripening. PLoS ONE 16(11):e0260004
61. Li M et al (2022) Integrating transcriptomic and metabolomic analysis in roots of wild soybean seedlings in response to low-phosphorus stress. Front Plant Sci 13:1006806
62. Sugiyama A (2019) The soybean rhizosphere: metabolites, microbes, and beyond—a review. J Adv Res 19:67–73
63. Bont Z et al (2020) Heritable variation in root secondary metabolites is associated with recent climate. J Ecol 108(6):2611–2624
64. Chen Q et al (2016) Arogenate dehydratase isoforms differentially regulate anthocyanin biosynthesis in *Arabidopsis thaliana*. Mol Plant 9(12):1609–1619
65. Sun XQ, Mao ZX, Fu H et al (2014) Fatty acid characteristics of forage and its influence factors. Pratacult Sci 31(9):1774–1780
66. Cui D, Liu Y, Yu H et al (2021) Geographical traceability of soybean based on elemental fingerprinting and multivariate analysis. J Consum Prot Food Saf 16(4):323–331

Publisher's Note