scientific data

OPEN

DATA DESCRIPTOR

Global transcription and metabolic profiles of five tissues in pepper fruits

Chengsheng Gong, Guangjun Guo, Baogui Pan, Changzhou Gao, Xianwei Zhu, Jinbing Liu, Shubin Wang & Weiping Diao ✉

Studying the regulatory mechanisms in different tissues of pepper is crucial for understanding organ formation, growth, and development. However, relevant studies are far from sufficient. In the current study, the stipe, calyx, pericarp, placenta, and seed of ripe pepper were sampled, and metabolites were determined by the untargeted metabolomics method. Transcriptome sequencing was performed by Illumina NovaSeq 6000, and then a high-throughput data set was built. The results showed that a total of 4879 annotated metabolites were detected in 15 samples of the five tissues under positive and negative ion mode. A total of 110.66 Gb of clean data was obtained by transcriptome sequencing, the clean data of each sample reached 6.21 Gb, and a total of 35 336 annotated expression genes were obtained. Furthermore, validate the accuracy of the data by combining principal component analysis and other methods. In summary, this study provides valuable information for the genetic improvement and breeding of peppers, and it holds potential application value, particularly in enhancing the quality and nutritional value of pepper fruits.

Background & Summary

Pepper belongs to the Solanaceae *Capsicum* genus of crops[1], and is one kind of horticultural crop planted worldwide. It is worth noting that the pepper fruit can be eaten fresh and used as an important condiment, which is deeply loved by consumers. Meanwhile, the rich metabolic components in chili peppers offer direct or indirect benefits to human health and pharmaceutical production, among other fields[2]. As the market's demand for high-quality vegetables increases, opportunities and challenges arise for the vegetable industry. Interestingly, integrating transcriptomic and metabolomic data offers new insights into understanding the growth and development, metabolic pathways, and gene regulation in pepper.

Metabolomics is a popular subject that studies metabolites[3], it has been widely studied in plants in recent years. Notably, metabolites are the basis for maintaining life activities, and are closely related to the growth and development of plants[4–6]. For instance, sugars, organic acids, and nucleotides provide essential energy for plant growth and development through a complex metabolic regulatory network[7,8]. In parallel, metabolites are the important link between genotype and phenotype, for example, carotenoids are metabolites associated with watermelon fruit color[9], cucurbitin is the key metabolite that causes the bitter taste of bitter gourd[10], and capsaicin is an important metabolite of pepper[11]. It is worth noting that in crops such as cabbage (*Brassica oleracea* var. *capitata* Linnaeus), sunflower (*Abelmoschus manihot* L. Medicus), and tomato (*Solanum lycopersicum* L.), critical research progress has been made in the study of metabolomics in different tissues, accelerating the synthesis pathway of important metabolic substances[12–15]. Importantly, with the continuous development of metabolomics detection technology, high-throughput data has become a display[16–18], such as untargeted metabolomics method to detect and analyze all metabolites that can be detected in the sample without bias.

The rapid development of transcriptome data has accelerated the mining of key genes for important traits and the in-depth study of key metabolic pathways. For example, by conducted transcriptome sequencing on 23 cucumber tissues, and combined with functional studies, found that *TERPENE SYNTHASE11* (*TPS11*)/*TPS14*, *TPS01*, and *TPS15* were responsible for the production of volatile terpenes in root, flower and fruit tissues of cucumber, respectively[19]. And through transcriptome determination of different tissues, important progress has been made in the mining of key genes for the formation of important traits in watermelon, asparagus, and

Jiangsu Key Laboratory for Horticultural Crop Genetic Improvement, Institute of Vegetable Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, 210014, China. ✉e-mail: diaowp_2000@163.com

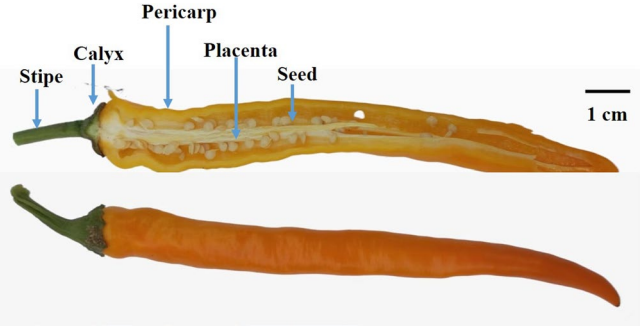

Fig. 1 Appearance diagram of different tissue parts of pepper fruit.

other crops[20,21]. Unfortunately, research on transcriptional regulation of different tissues in peppers has mainly focused on nutrient organs such as roots, stems, and leaves[22], while research on transcriptional regulation of reproductive organs such as fruits is still insufficient.

Transcriptome studies have made some progress in the analysis of the formation of carotenoids and other traits in pepper[23]. However, metabolomics studies mostly focus on a small number of metabolites such as capsaicin[24], and there are no systematic transcriptome and metabolomics studies on different fruit tissues. In the current study, the stipe, calyx, pericarp, placenta, and seed of the pepper DC324 were sampled as experimental materials (Fig. 1), untargeted metabolomics methods were used to detect metabolites, second-generation sequencing technique was used for transcriptome sequencing to obtain data sets, and further validate the accuracy of these data. In summary, this study has yielded a rich dataset, demonstrating a certain level of innovation in metabolism and transcriptional regulation across multiple tissues of pepper fruit. It provides significant references for future research on genetic mechanisms, genetic improvement, and breeding efforts.

Methods

Germplasm source and cultivation management of pepper. The tested pepper DC324 belongs to *Capsicu mannuum* L., which is an inbred line germplasm resource of the line pepper, and was independently bred by the Vegetable Research Institute of Jiangsu Academy of Agricultural Sciences. In April of 2023, pepper seedlings were planted in the greenhouse of Luhe Animal Science base of Jiangsu Academy of Agricultural Sciences (Nanjing, Jiangsu, East longitude: 118.83, north latitude: 32.35) through seedling cultivation and transplanting. 30 seedlings were planted with 50 cm plant spacing and 50 cm row spacing. Pollination was carried out by self-pollination, and the date of pollination was marked with a label. The same fertilization, irrigation, and other cultivation management measures were adopted in the whole growth cycle, for example, the main fertilizers include pure N 5.0 kg, P_2O_5 5-6 kg, K_2O 12–15 kg, and include about 15 irrigation cycles.

Sample collection of test materials. At the ripening stage of the pepper fruit (the color of the peel completely turned yellow), samples were collected from peppers with relatively consistent growth at 55 days after pollination. Three peppers from the same plant were mixed as a biological replicate, and peppers from three plants were taken as three biological replicates. During the sample collection, the stipe and calyx were separated first, and then the pepper fruit was cut longitudinally, and the pericarp, placenta, and seeds were sampled respectively. Each tissue was mixed and put into a 50 mL conical tube, then, put into liquid nitrogen for quick freezing, and transferred to the refrigerator at −80 °C until use.

Metabolite determination using untargeted metabolomics method. The sample to be tested was fully ground into powder in the grinder, 50 mg of the freeze-dried sample was weighed for metabolite determination, 1000 μL of the extraction solution containing the internal standard was added (methanol acetonitrile-water volume ratio = 2:2:1, internal standard concentration 20 mg/L), which was used to correct for various biases that may occur during the analysis, and vortex mixed for 30 seconds; then add the steel ball to the 45 Hz grinder for 10 min, ultrasonic 10 min; the sample to be tested was obtained after filtration. When detecting metabolites, metabolite determination was performed based on the LC-MS system, which mainly consists of Waters Acquity I-Class PLUS ultra-high performance liquid tandem and Waters Xevo G2-XS QTof high-resolution mass spectrometer. The Waters Acquity UPLC HSS T3 column (1.8 um 2.1 × 100 mm) was used as the chromatographic column. Positive and negative ion modes were used to determine the metabolites. Mobile phase A: 0.1% formic acid aqueous solution; Mobile phase B: 0.1% acetonitrile formate. The mobile phase conditions of liquid chromatography were as follows: the flow rate was 400 μL/min, 0.0 min: 98% flow A, 2% flow B; 0.25 min: 98% flow A, 2% flow B, 10.0 min: 2% flow A, 98% flow B; 13.0 min: 2% flow A, 98% flow B; 13.1 min: 98% flow A, 2% flow B; 15.0 min: 98% flow A, 2% flow B. MSe mode controlled by acquisition software (MassLynx V4.2, Waters) was used for primary and secondary mass spectrum data acquisition. ESI ion source parameters are as follows: Capillary voltage: 2500 V (positive ion mode) or −2000 V (negative ion mode); Cone hole voltage: 30 V; Ion source temperature: 100 °C; Desolvent temperature 500 °C; Air flow rate: 50 L/h; Desolvent gas flow rate: 800 L/h; Plastic-nucleus ratio (m/z) collection range 50–1200. In the qualitative and quantitative analysis of metabolites, the original data collected by MassLynx V4.2 were processed by the Progenesis QI software for peak extraction, peak alignment, and other data, and the metabolites were identified based on the Progenesis QI software online

METLIN database. Then, based on the results of the total score, MS2 score, and mass Error (ppm), the metabolites were qualitatively determined[25]. The final metabolite types were determined by the artificial deletion of duplicate data. KEGG (http://www.genome.jp/kegg/), HMDB (https://hmdb.ca/) and Lipidmaps (https://lipidmaps.org/) were the main databases for metabolome data comparison.

Transcriptome sequencing and gene expression level analysis. Transcriptome sequencing is a mature technology, with high throughput, high resolution, wide species availability, high sensitivity, wide detection range, and dynamic monitoring of gene expression levels. Total RNA was extracted from plants using RNA prep Pure Plant Kit (Tiangen, Beijing, China). The purity and concentration of RNA were measured using the NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE), and the integrity of RNA was measured using the Agient2100/LabChip GX (Agilent Technologies, CA, USA) to obtain high-quality RNA. Sequencing libraries were generated using the Hieff NGS Ultima Dual-mode mRNA Library Prep Kit for Illumina (Yeasen Biotechnology (Shanghai) Co., Ltd.). Index was added to the sequence for each sample. The libraries were sequenced on the Illumina NovaSeq 6000 platform (San Diego) to generate a 150 bp double-terminal sequence. Raw data in Fastq format was first processed by Perl scripts, while referring to Ewing, *et al.*[26] 's method to calculate base Q-score. And the Zunla pepper genome[27] was used as the reference genome for follow-up analysis. After the original quality control through FastQC (v0.11; https://github.com/s-andrews/FastQC), the valid data were then compared by HISAT2 (2.0.4)[28] and SAMTools (v1.12)[29] to the reference genome sequence. Transcriptome splicing was combined using StringTie (v2.2.1)[30]. And the software was used to identify new transcripts to mine for new genes. Then, the FPKM (Fragments Per Kilobase of exon model per Million mapped fragments) was used to measure the level of gene expression, and the gene enrichment analysis was realized by GO (Gene Ontology).

Data statistics and graph rendering. EXCEL2023 was utilized for data analysis and statistical processing, while R software was primarily employed for graph generation. Specifically, the PCA prcomp, pheatmap package, and corrplot package were used to create principal component analysis figures, heatmaps, and correlation analysis figures.

Data Records

In the current study, the raw data for metabolomics can be obtained in Metabolights, with the ID MTBLS10002 (MetaboLights MTBLS)[31]. The original transcriptome sequencing data was stored in NCBI Sequence Read Archive (SRA) accession (No. PRJNA1101187)[32], and the SRR numbers for 15 samples range from SRR28741309 to SRR28741323. Meanwhile, transcriptome and metabolome expression data have been uploaded to the figshare database for easy download and reference[33] (sdata Figshare repository https://doi.org/10.6084/m9.figshare.27054586).

Technical Validation

Biological repetition. In the current study, fruit samples from different pepper plants were collected for three biological replicates to improve the reliability of the assay results.

Verification of metabolites. Firstly, in terms of qualitative and quantitative analysis of metabolites, the original data collected by MassLynx V4.2 were processed by the Progenesis QI software for peak extraction, peak alignment and other data processing operations. The identification was carried out based on the online METLIN database of Progenesis QI software, the public database and the self-built database of Biomarker, and the theoretical fragment identification was also carried out. The mass number deviation of the parent ion is 100 ppm, and the mass number deviation of the fragmention was less than 50 ppm; the samples were measured in the same batch to ensure the reliability of the determination results, and the relative accumulation amount was used as an important indicator to measure the difference between different samples. And a total of 4 879 metabolites with annotated information were obtained.

Secondly, when correlation analysis and principal component analysis were performed on metabolome data, it was found that different tissues were significantly separated, and the two principal components explained 31.69% and 26.78% of the variation, respectively (Fig. 2a). Meanwhile, there was a strong correlation between the three biological replicates of the same tissue (Fig. 2b). The results of the KEGG enrichment analysis indicate that metabolites are significantly enriched in pathways such as "Biosynthesis of Other Secondary Metabolites", suggesting that different tissues are engaged in robust metabolic activities (Fig. 2c). Thus, by conducting thorough analyses, we've amassed a comprehensive dataset of metabolites. These results help us understand the metabolic basis for the formation of functional specificity in different tissues. This variation was considered to be closely related to tissue functional specificity.

Validation of transcriptome data. Among the 15 tested samples, a total of 110.66 Gb of clean data was obtained, the percentage of Q30 bases in each sample was no less than 93.90%, and the comparison efficiency of sequence alignment with reference genome ranged from 91.08% to 93.89% (Table 1). Through the distribution of Mapped Reads across various mRNA transcripts, it was found that gene expression is relatively consistent, with no evident degradation of the mRNA present (Fig. 3a). Therefore, the amount of data obtained was sufficient and can meet the requirements of subsequent analysis.

A total of 35 336 genes with gene annotation information were discovered. After further analysis of the distribution characteristics of FPKM data, it was found that the FPKM value of coding genes ranged from 10^{-2} to 10^{4}, and the expression levels of most genes were concentrated between 0.1 and 10 (Fig. 3b-c). Meanwhile, through the enrichment analysis of Gene Ontology (GO) for all genes, it was found that a greater number of expressed genes are enriched in the 'Metabolic Processes' category (Fig. 3d). This suggests that the expressed genes play

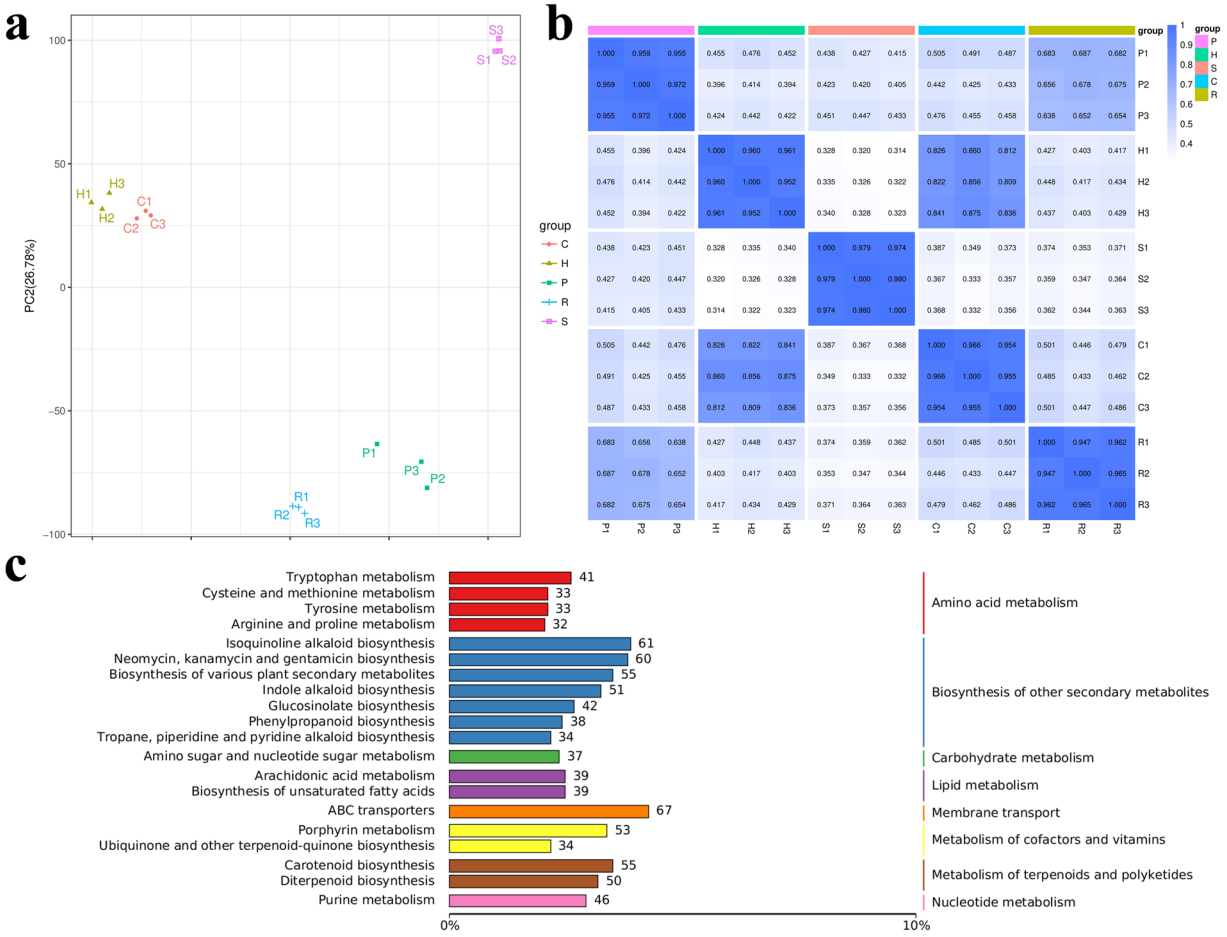

Fig. 2 Correlation analysis diagrams and enrichment analysis diagrams of all metabolites. (**a**) Principal component analysis diagram of metabolites (**b**) Correlation heat map of metabolites between samples to be tested. (**c**) All metabolites enrichment analysis. H, C, R, P, and S stand for stipe, calyx, pericarp, placenta, and seed respectively. −1, −2, −3 represent three biological repeats.

Samples	Clean reads	Clean bases	GC Content	% ≥ Q30	Total Reads	Mapped Reads
C-1	22,551,490	6,747,127,828	42.46%	93.90%	45,102,980	42,348,842 (93.89%)
C-2	28,945,693	8,669,540,148	42.10%	94.73%	57,891,386	52,932,560 (91.43%)
C-3	28,158,506	8,432,974,448	42.19%	95.74%	56,317,012	51,959,069 (92.26%)
H-1	29,372,823	8,797,201,222	42.28%	94.87%	58,745,646	53,982,730 (91.89%)
H-2	24,035,322	7,198,486,554	42.35%	94.94%	48,070,644	44,170,853 (91.89%)
H-3	23,338,645	6,989,308,336	42.04%	94.59%	46,677,290	42,628,846 (91.33%)
P-1	20,751,408	6,214,208,002	41.84%	94.74%	41,502,816	37,974,734 (91.50%)
P-2	23,509,655	7,040,829,242	42.30%	95.19%	47,019,310	43,541,038 (92.60%)
P-3	22,860,560	6,846,753,962	42.08%	95.26%	45,721,120	42,320,898 (92.56%)
R-1	26,699,719	7,997,842,696	42.09%	94.82%	53,399,438	49,073,607 (91.90%)
R-2	26,022,981	7,794,024,400	42.08%	94.99%	52,045,962	48,051,203 (92.32%)
R-3	23,580,515	7,063,388,502	42.26%	95.08%	47,161,030	43,651,699 (92.56%)
S-1	23,176,819	6,941,776,842	42.23%	95.66%	46,353,638	42,364,889 (91.39%)
S-2	24,843,070	7,440,962,514	42.08%	95.60%	49,686,140	45,675,667 (91.93%)
S-3	21,644,363	6,481,613,732	42.13%	95.12%	43,288,726	39,425,635 (91.08%)

Table 1. Transcriptome sequencing data statistics table. **Note:** H, C, R, P, and S stand for stipe, calyx, pericarp, placenta, and seed respectively. −1, −2, −3 represent three biological repeats.

a crucial regulatory role in aspects such as metabolite synthesis. Principal component analysis of all expressed genes showed that the first two principal components were 33.12% and 22.54%, respectively (Fig. 4a). The results of three biological repetitive clusterings showed that the same tissue sites had good repeatability (Fig. 4b), and

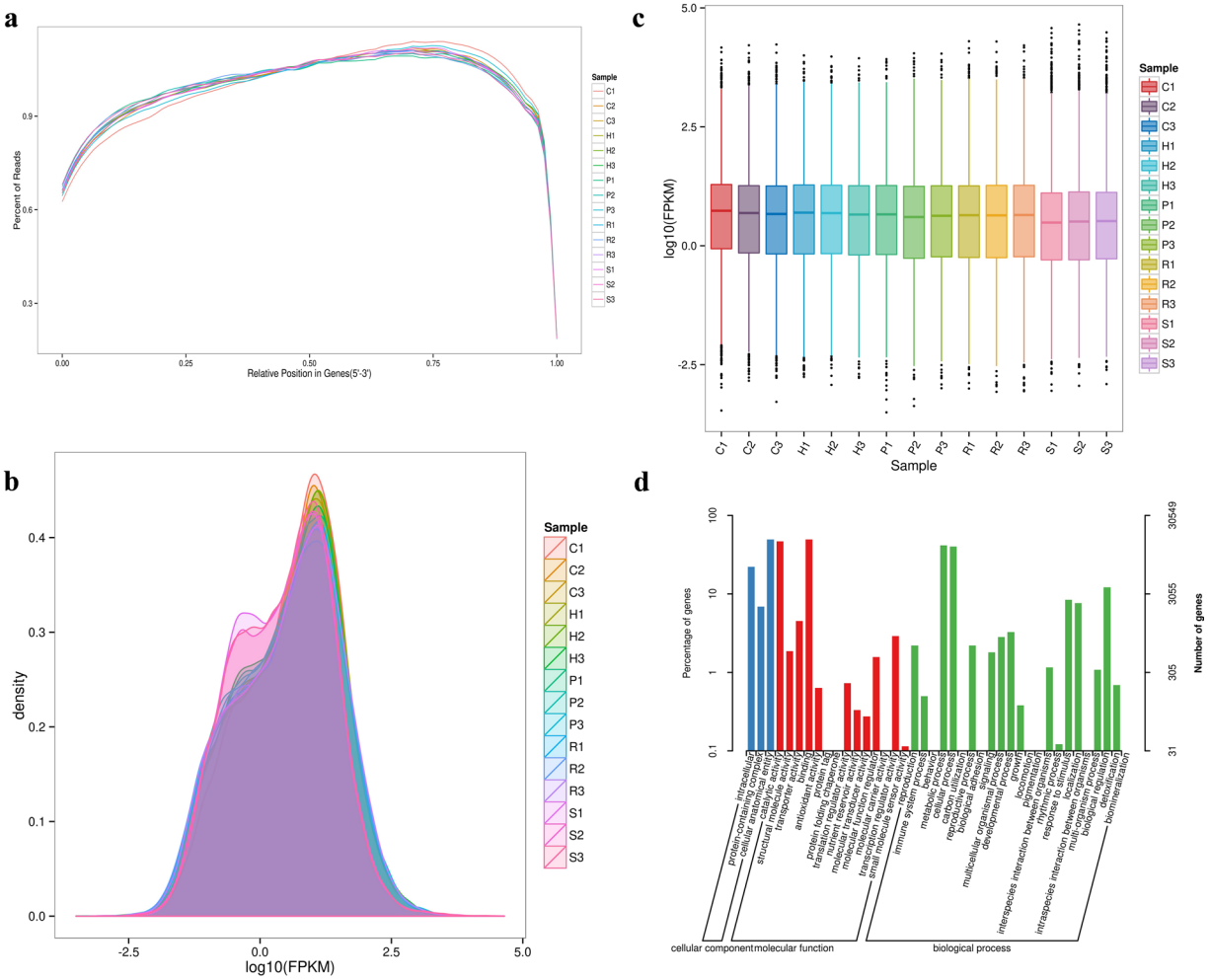

Fig. 3 Quality control and statistical chart for transcriptome data across 15 samples of five tissues. (**a**) Distribution chart of Mapped Reads on mRNA. (**b**) Comparison of FPKM density distribution; (**c**) FPKM container line diagram. (**d**) Enrichment analysis chart for all genes using Gene Ontology (GO). Different colors in the picture represent different samples. H, C, R, P, and S stand for stipe, calyx, pericarp, placenta, and seed respectively.

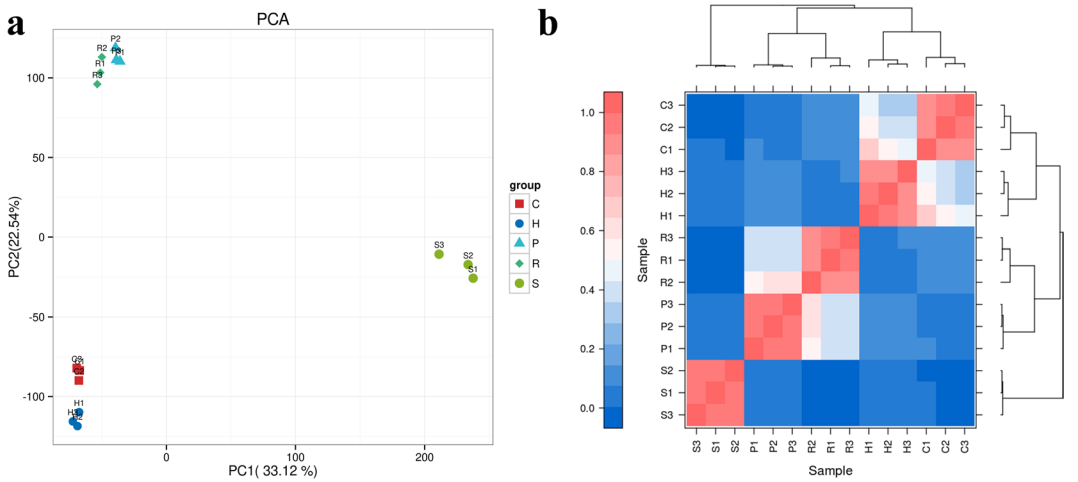

Fig. 4 Correlation analysis diagram of all genes (**a**) Principal component analysis diagram of genes (**b**) Correlation heat map of genes between samples to be tested. H, C, R, P, and S stand for stipe, calyx, pericarp, placenta, and seed, respectively. −1, −2, and −3 represent three biological repeats.

there were significant differences in different tissue sites. Rich transcriptomic data offer a robust foundation for deciphering metabolite formation, tissue-specific expression, and related studies.

Code availability

In the current study, data processing and analysis were mainly carried out through R software (3.6.1), related plug-ins, and data scripts, mainly obtained from publicly published data, and all the code was publicly available. For PCA analysis, PCA prcomp (R base function) 3.6.1 was used, and the scale method was uv scaling. For heatmap analysis, the pheatmap package (1.0.2) of R software was used, and the scale mode was UV scaling. The corrupt package (0.73) of R software was used for the correlation heat map, and the significance threshold was P value ≤ 0.05. We did not use additional custom code in this study.

Received: 31 May 2024; Accepted: 27 September 2024;
Published online: 15 October 2024

References

1. Carrizo García, C. *et al*. Phylogenetic relationships, diversification and expansion of chili peppers (Capsicum, Solanaceae). *Ann Bot* **118**, 35–51, https://doi.org/10.1093/aob/mcw079 (2016).
2. Liu, F. *et al*. Genomes of cultivated and wild Capsicum species provide insights into pepper domestication and population differentiation. *Nature Communications* **14**, 5487, https://doi.org/10.1038/s41467-023-41251-4 (2023).
3. Shen, S., Zhan, C., Yang, C., Fernie, A. R. & Luo, J. Metabolomics-centered mining of plant metabolic diversity and function: Past decade and future perspectives. *Mol Plant* **16**, 43–63, https://doi.org/10.1016/j.molp.2022.09.007 (2023).
4. Liu, Y.-L. *et al*. Morphological, physiochemical, and transcriptome analysis and CaEXP4 identification during pepper (Capsicum annuum L.) fruit cracking. *Sci Hortic-Amsterdam* **297**, 110982, https://doi.org/10.1016/j.scienta.2022.110982 (2022).
5. Lee, S. Y. *et al*. A mutation in Zeaxanthin epoxidase contributes to orange coloration and alters carotenoid contents in pepper fruit (Capsicum annuum). *Plant J* **106**, 1692–1707, https://doi.org/10.1111/tpj.15264 (2021).
6. Byun, J. *et al*. Identification of CaAN3 as a fruit-specific regulator of anthocyanin biosynthesis in pepper (Capsicum annuum). *Theor Appl Genet* **135**, 2197–2211, https://doi.org/10.1007/s00122-022-04106-y (2022).
7. Maeda, H. A. Harnessing evolutionary diversification of primary metabolism for plant synthetic biology. *J Biol Chem* **294**, 16549–16566, https://doi.org/10.1074/jbc.REV119.006132 (2019).
8. Ahmad, I. *et al*. Diversity and expression analysis of ZIP transporters and associated metabolites under zinc and iron stress in Capsicum. *Plant Physiol Biochem* **196**, 415–430, https://doi.org/10.1016/j.plaphy.2023.01.060 (2023).
9. Yuan, P. L. *et al*. Transcriptome regulation of carotenoids in five flesh-colored watermelons. *Bmc Plant Biology* **21**, 203–203, https://doi.org/10.1186/s12870-021-02965-z (2021). doi:ARTN 203.
10. Zhou, Y. *et al*. Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nature plants* **2**, 16183–16183, https://doi.org/10.1038/nplants.2016.183 (2016).
11. Li, R. *et al*. Capsaicin Attenuates Oleic Acid-Induced Lipid Accumulation via the Regulation of Circadian Clock Genes in HepG2 Cells. *J Agric Food Chem* **70**, 794–803, https://doi.org/10.1021/acs.jafc.1c06437 (2022).
12. Wei, L. *et al*. Spatio-temporal transcriptome profiling and subgenome analysis in Brassica napus. *Plant J* **111**, 1123–1138, https://doi.org/10.1111/tpj.15881 (2022).
13. Zhou, Y. *et al*. Combined Metabolome and Transcriptome Analyses Reveal the Flavonoids Changes and Biosynthesis Mechanisms in Different Organs of Hibiseu manihot L. *Front Plant Sci* **13**, 817378, https://doi.org/10.3389/fpls.2022.817378 (2022).
14. Tohge, T. *et al*. Exploiting Natural Variation in Tomato to Define Pathway Structure and Metabolic Regulation of Fruit Polyphenolics in the Lycopersicum Complex. *Mol Plant* **13**, 1027–1046, https://doi.org/10.1016/j.molp.2020.04.004 (2020).
15. Li, Y. *et al*. MicroTom Metabolic Network: Rewiring Tomato Metabolic Regulatory Network throughout the Growth Cycle. *Mol Plant* **13**, 1203–1218, https://doi.org/10.1016/j.molp.2020.06.005 (2020).
16. Yuan, H. L. *et al*. Development of a widely targeted volatilomics method for profiling volatilomes in plants. *Molecular Plant* **15**, 189–202, https://doi.org/10.1016/j.molp.2021.09.003 (2022).
17. He, J. M. *et al*. A Sensitive and Wide Coverage Ambient Mass Spectrometry Imaging Method for Functional Metabolites Based Molecular Histology. *Adv Sci* **5**, 1800250, https://doi.org/10.1002/advs.201800250 (2018).
18. 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. *Molecular Plant* **6**, 1769–1780, https://doi.org/10.1093/mp/sst080 (2013).
19. Wei, G. *et al*. Integrative Analyses of Nontargeted Volatile Profiling and Transcriptome Data Provide Molecular Insight into VOC Diversity in Cucumber Plants (Cucumis sativus). *Plant Physiol* **172**, 603–618, https://doi.org/10.1104/pp.16.01051 (2016).
20. Srivastava, P. L., Shukla, A. & Kalunke, R. M. Comprehensive metabolic and transcriptomic profiling of various tissues provide insights for saponin biosynthesis in the medicinally important Asparagus racemosus. *Sci Rep* **8**, 9098, https://doi.org/10.1038/s41598-018-27440-y (2018).
21. Gong, C. S. *et al*. An integrated transcriptome and metabolome approach reveals the accumulation of taste-related metabolites and gene regulatory networks during watermelon fruit development. *Planta* **254**, 35, https://doi.org/10.1007/s00425-021-03680-7 (2021).
22. Liao, Y. *et al*. The 3D architecture of the pepper genome and its relationship to function and evolution. *Nat Commun* **13**, 3479, https://doi.org/10.1038/s41467-022-31112-x (2022).
23. Tang, Y. *et al*. Identification of carotenoids and candidate genes shaping high pigment chili pepper variety. *Sci Hortic-Amsterdam* **327**, 112799, https://doi.org/10.1016/j.scienta.2023.112799 (2024).
24. Nakaniwa, R. *et al*. Biochemical Aspects of Putative Aminotransferase Responsible for Converting Vanillin to Vanillylamine in the Capsaicinoid Biosynthesis Pathway in Capsicum Plants. *J Agric Food Chem* **72**, 559–565, https://doi.org/10.1021/acs.jafc.3c07369 (2024).
25. Wang, J. *et al*. Serum metabolomics for early diagnosis of esophageal squamous cell carcinoma by UHPLC-QTOF/MS. *Metabolomics* **12**, 116, https://doi.org/10.1007/s11306-016-1050-5 (2016).
26. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175–185, https://doi.org/10.1101/gr.8.3.175 (1998).
27. Qin, C. *et al*. Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proc Natl Acad Sci USA* **111**, 5135–5140, https://doi.org/10.1073/pnas.1400975111 (2014).
28. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
29. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).

30. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
31. Gong, C. *et al*. MetaboLights, MTBLS10002, https://www.ebi.ac.uk/metabolights/MTBLS10002.
32. *NCBI Sequence Read Archive* https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA1101187 (2024).
33. Figshare. https://doi.org/10.6084/m9.figshare.27054586.

Acknowledgements

Thanks to the following grants for support of this study: the Natural Science Foundation of Jiangsu Province, China (BK20230751), Jiangsu Key Laboratory for Horticultural Crop Genetic Improvement (ZD2022003), Jiangsu agricultural science and technology innovation fund [CX(24)3019], Jiangsu key R & D plan (BE2023349). Statement of Interest: In the current study, the authors declare that they are bound by confidentiality agreements that prevent them from disclosing their competing interests in this work. At the same time, nothing in this manuscript has any potential conflict of interest with any institution, etc.

Author contributions

Chengsheng Gong: Data curation, Writing – original draft. Guangjun Guo: Formal analysis, Validation, Visualization. Baogui Pan: Formal analysis, Validation, Visualization, Writing – review & editing. Changzhou Gao: Project administration, Resources, Software, Supervision. Xianwei Zhu: Investigation, Methodology, Resources, Software, Supervision. Jinbing Liu: Project administration, Resources, Software, Supervision. Shubin Wang: Project administration, Resources, Software, Supervision. Weiping Diao: Conceptualization, Funding acquisition.

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.

Additional information

Correspondence and requests for materials should be addressed to W.D.